Goodness-of-fit testing in high-dimensional generalized linear models

Jana Janková*, Rajen D. Shah*, Peter Bühlmann† and Richard J. Samworth*
*University of Cambridge and †ETH Zürich

March 11, 2020

Abstract

We propose a family of tests to assess the goodness-of-fit of a high-dimensional generalized linear model. Our framework is flexible and may be used to construct an omnibus test or directed against testing specific non-linearities and interaction effects, or for testing the significance of groups of variables. The methodology is based on extracting left-over signal in the residuals from an initial fit of a generalized linear model. This can be achieved by predicting this signal from the residuals using modern powerful regression or machine learning methods such as random forests or boosted trees. Under the null hypothesis that the generalized linear model is correct, no signal is left in the residuals and our test statistic has a Gaussian limiting distribution, translating to asymptotic control of type I error. Under a local alternative, we establish a guarantee on the power of the test. We illustrate the effectiveness of the methodology on simulated and real data examples by testing goodness-of-fit in logistic regression models. Software implementing the methodology is available in the R package `GRPtests` (Janková et al., 2019).

1 Introduction

In recent years, there has been substantial progress in developing methodology for estimation in generalized linear models (GLMs) in high-dimensional settings, where the number of covariates in the model may be much larger than the number of observations. A standard technique for estimation is the Lasso for generalized linear models (Park and Hastie, 2007), which has a fast implementation in the R package `glmnet` (Friedman et al., 2010) and is widely used. The Lasso enjoys good empirical and theoretical properties for estimation and variable selection, provided that we are searching for a sparse approximation to the regression coefficients in the generalized linear model.

Once a generalized linear model has been fitted to the high-dimensional data, it is important to assess the quality of the fit. Literature on testing goodness-of-fit in low-dimensional settings is extensive: we refer to Section 1.2 below for an overview. However, the methods typically rely on properties that only hold in low-dimensional settings such as asymptotic linearity and normality of the maximum likelihood estimator, for example. These may fail to hold with an increasing number of covariates in the model; as a consequence it is typically not possible to extend these approaches

in an obvious way to the high-dimensional setting. This motivates us to develop a new method that may be used for detecting misspecification in the fit of a (potentially high-dimensional) generalized linear model.

To fix ideas, suppose we have data $(x_i, Y_i)_{i=1}^n$ formed of feature vectors $x_i \in \mathbb{R}^p$ and univariate responses $Y_i \in \mathcal{Y} \subseteq \mathbb{R}$. Let us write $X = [x_1, \ldots, x_n]^T = [X_1, \ldots, X_p] = (X_{ij})_{1 \le i \le n, 1 \le j \le p}$ for the $n \times p$ design matrix and $Y = (Y_1, \ldots, Y_n)^T$ for the vector of responses. Specifically, consider the setting where the Y_i are independent conditional on X, and where the conditional distribution of Y_i only depends on X through x_i. Moreover, suppose that the conditional expectation and conditional variance have structures of the form

$$m_0(x) := \mathbb{E}(Y_i | x_i = x) = \mu(x^T \beta_0) \quad \text{and} \quad \mathrm{Var}(Y_i | x_i = x) = V(\mu(x^T \beta_0))$$

for some unknown $\beta_0 \in \mathbb{R}^p$, where $\mu(\cdot)$ is a known inverse link function and $V(\cdot)$ is also known. The main examples we have in mind are generalized linear models (McCullagh and Nelder, 1989), including, for example, logistic regression, probit regression or Poisson log-linear models, but our framework also allows quasi-GLMs, and other settings where the variance depends on the features only through the means as imposed in the condition above.

We will focus on the detection of misspecification in the conditional mean function. In a low-dimensional setting, we understand that the model is misspecified in the conditional mean when there does not exist a $\beta_0 \in \mathbb{R}^p$ such that $m_0(x) = \mu(x^T \beta_0)$. In a high-dimensional setting where $p \ge n$, this concept becomes more complicated at first sight; for example, with fixed design points $x_1, \ldots, x_n$, there always exists $\beta_0 \in \mathbb{R}^p$ such that $m_0(x_i) = \mu(x_i^T \beta_0)$ for all $i = 1, \ldots, n$, meaning that the model can never be misspecified. However, in a high-dimensional setting, it is impossible to estimate β_0 consistently without additional structural assumptions. An assumption that is often used, and which we adopt in this paper, is sparsity of the model. Therefore, we address the question of whether a *sparse* model fits well to the observations, or whether a (sparse) non-linear model is more appropriate. If we restrict ourselves to sparse models, then misspecification can happen in the same way as in low-dimensional settings, even for fixed design. Some of the most important types of misspecification that are of interest in applications are missing nonlinear terms such as quadratic effects or interaction terms.

1.1 Overview of our contributions

We now briefly outline our strategy for goodness-of-fit testing; a more detailed description is given in Section 2. Let $\hat{\beta}$ be an estimate of β_0 derived from a Lasso-penalised generalized linear model (GLM Lasso) fit. Our starting point is the vector R of Pearson residuals, with i-th coordinate

$$R_i := \frac{Y_i - \mu(x_i^T \hat{\beta})}{\sqrt{V(\mu(x_i^T \hat{\beta}))}}, \quad i = 1, \ldots, n.$$

Now consider taking as a test statistic the scalar product $w^T R$, for some (fixed) unit vector $w \in \mathbb{R}^n$. If the generalized linear model were correct, then $w^T R$ would be approximately an average of zero-mean random variables, and under reasonable conditions, should converge to a centred Gaussian

random variable. On the other hand, if the model were misspecified, the residuals would contain some signal, and were w to be positively correlated with this signal, the lack of fit should be exposed by the test statistic taking a large value.

In the alternative setting, the signal in the residuals may be picked up by more flexible regression methods, such as random forests (Breiman, 2001) or boosted trees (Freund and Schapire, 1996; Chen and Guestrin, 2016). However using such flexible regressions to inform the choice of w directly would make w strongly dependent on R even under the null; as such calibration of the resulting test statistic would be problematic. Our approach therefore is to construct w based on an independent auxiliary dataset (X_A, Y_A) (e.g. derived through sample splitting) in the following way. We first perform a GLM Lasso fit on the auxiliary dataset to obtain an additional set of residuals. Regressing these residuals back on to the explanatory variables X_A using a nonlinear regression method, we obtain an estimated regression function $\tilde{f} : \mathbb{R}^p \to \mathbb{R}$ that aims to predict the signal in the residuals; we refer to the n-fold concatenation of such an $\tilde{f}$ as a *residual prediction function* $\hat{f} : \mathbb{R}^{n \times p} \to \mathbb{R}^n$. We may then choose w proportional to $\hat{f}(X)$ to give a direction w independent of Y.

One important issue that arises in the high-dimensional setting is that although components of R are close to zero-mean under the null, their bias can drive a substantial shift in the mean of $w^T R$. To prevent this, we replace w with the residuals from a particular weighted (square-root) Lasso regression of w on to X. This final step ensures that w is almost orthogonal to the bias in the residuals and as a consequence, the limiting distribution under the null is a centred Gaussian. A notable feature of our construction is that the asymptotic null distribution is essentially invariant to the residual prediction method used. This can therefore be as flexible as needed to detect the type of mean misspecification we would like to uncover.

We provide a software implementation of our methodology in the R package `GRPtests` (Janková et al., 2019).

1.2 Related literature

High dimensions. Our work is related to that of Shah and Bühlmann (2018) who study goodness-of-fit tests for the linear model. They consider test statistics based on a proxy for the prediction error of a flexible regression method applied to the scaled residuals following a square-root Lasso fit to the data. It is shown that when a Gaussian linear model holds, these residuals depend only weakly on the unknown regression coefficients, motivating calibration of the tests via a parametric bootstrap. As there is no analogue of this result for other generalized linear models, it does not seem possible to extend this approach to our more general setting. Our methodology shares the idea of 'predicting' the residuals but, even when we specialize our approach to the Gaussian linear model, differs substantially in the construction of test statistics and the form of calibration.

In recent years, there has been much work on inference and testing in high-dimensional generalized linear models, particularly for the linear model. The work on significance testing includes flexible approaches based on (multiple) sample-splitting (Wasserman and Roeder, 2009; Meinshausen et al., 2009; Meinshausen and Bühlmann, 2010; Shah and Samworth, 2013) which may be combined with other methods. Another line of work, initiated by Zhang and Zhang (2014), proposes a

method of de-biasing the Lasso that can be used for testing significance of variables in the linear regression. van de Geer et al. (2014) extend the methodology to generalized linear models; further developments include Javanmard and Montanari (2014), Dezeure et al. (2017) and Yu et al. (2020); see also Belloni et al. (2014). General frameworks for testing low-dimensional hypotheses about the parameter can be based for example on Neyman orthogonality conditions (Chernozhukov et al., 2015, 2018) or on a profile likelihood testing framework (Ning et al., 2017). In recent work, Zhu and Bradic (2017) propose a method for testing more general hypotheses about the parameter vector, such as the sparsity level of the model parameter and minimum signal strength. Javanmard and Lee (2017) suggest a procedure to test similar hypotheses about the parameter in linear or logistic regression.

Low dimensions. There are numerous methods for testing goodness-of-fit of a model in low-dimensional settings, especially for the case of logistic regression, which is one of the focuses of this work. The most standard tests are residual deviance and Pearson's chi-squared tests; however, they behave unsatisfactorily if the data contain only a small number of observations for each pattern of covariate values. There have been a number of strategies to circumvent this difficulty, mainly based on grouping strategies, residual smoothing or modifications of Pearson's chi-squared test.

Hosmer and Lemeshow (1980) proposed two methods of grouping based on ranked estimated logistic probabilities that form groups of equal numbers of subjects. The disadvantage of these tests (as noted in Le Cessie and Van Houwelingen (1991)) is that as they are based on a grouping strategy in the space of responses, they lack power to detect departures from the model in regions of the covariate space that yield the same estimated probabilities. For example, a model with a quadratic term may have very different covariate values with the same estimated probability. Tsiatis (1980) circumvents the difficulties faced by Hosmer–Lemeshow tests using a grouping strategy in the covariate space. However, different partitions of the space of covariates may still lead to substantially different conclusions.

Le Cessie and Van Houwelingen (1991) introduced a test based on residual smoothing using nonparametric kernel methods. Smoothed residuals replace each residual with a weighted average of itself and other residuals that are close in the covariate space. If residuals close to each other are strongly correlated, smoothing does not affect the magnitude of the residuals strongly, while if they are not correlated smoothing will shrink the residuals towards zero. Su and Wei (1991) proposed a goodness-of-fit test for the generalized linear model based on a cumulative sum of residuals, which was later adapted by Lin et al. (2002) and a weighted version was proposed in Hosmer and Hjort (2002). Another approach based on modifications of Pearson's chi-squared test was studied in Osius and Rojek (1992) and Farrington (1996) who derived a large-sample normal approximation for Pearson's chi-squared test statistic.

1.3 Organization and notation

The rest of the paper is organised as follows. In Section 2 we motivate and present our goodness-of-fit testing methodology. In Section 3, we study its theoretical properties, providing guarantees on the type I error and power. In Section 4, we illustrate the empirical performance of the method on simulated and semi-real genomics data. A brief discussion is given in Section 5. Proofs are deferred

to Section 6 and Appendix A.

For a vector $x \in \mathbb{R}^d$, we let x_j denote its j-th entry and write $\|x\|_p := (\sum_{j=1}^d |x_j|^p)^{1/p}$ for $p \in \mathbb{N}$, $\|x\|_\infty := \max_{j=1,\dots,d} |x_j|$ and $\|x\|_0$ for the number of non-zero entries of x. For a matrix $A \in \mathbb{R}^{n\times p}$, we use the notation A_{ij} or $(A)_{ij}$ for its (i,j)-th entry, A_j to denote its j-th column and we let $\|A\|_\infty := \max_{i,j} |A_{ij}|$. Letting $G \subseteq \{1,\dots,p\}$, we denote by A_G the matrix containing only columns from A whose indices are in G, and by A_{-G} the columns of A whose indices are in the complement of G. We use $\Lambda_{\min}(A)$ and $\Lambda_{\max}(A)$ to denote the minimum and maximum eigenvalue of a square matrix A.

For sequences of random variables X_n, Y_n, we write $X_n = \mathcal{O}_P(Y_n)$ if X_n/Y_n is bounded in probability and $X_n = o_P(1)$ if X_n converges to zero in probability. We write $a \lesssim b$ to mean that there exists $C > 0$, which may depend on other quantities designated as constants in our assumptions, such that $a \leq Cb$. If $a \lesssim b$ and $b \lesssim a$, we write $a \asymp b$. Finally, for a function $f : \mathbb{R} \to \mathbb{R}$ and a vector $z = (z_1,\dots,z_n) \in \mathbb{R}^n$ we will use $f(z)$ to denote the coordinate-wise application of f to z, that is $f(z) = (f(z_1),\dots,f(z_n))$.

2 Methodology: Generalized Residual Prediction tests

As mentioned in Section 1.1, our Generalized Residual Prediction (GRP) testing methodology relies on an initial fit of the form

$$\hat{\beta} := \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \left\{ \frac{1}{n}\sum_{i=1}^n \rho(Y_i, x_i^T\beta) + \lambda\|\beta\|_1 \right\}.$$

In what follows, we refer to $\hat{\beta}$ as the GLM Lasso, though it is not essential that the loss function $\rho : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ is the negative log-likelihood obtained from a generalized linear model, and indeed this definition incorporates penalized quasi-likelihood estimators, amongst others. Our general framework for goodness-of-fit testing will also assume we have available an auxiliary dataset $(X_A, Y_A) \in \mathbb{R}^{n_A\times p} \times \mathcal{Y}^{n_A}$ independent of (X, Y). In the rest of the paper, we take $n_A = n$ for simplicity, although this is not needed for our procedures. Consider the Pearson-type residuals

$$R_i = \frac{Y_i - \mu(x_i^T\hat{\beta})}{\sqrt{V(\mu(x_i^T\tilde{\beta}))}}, \quad i = 1,\dots,n.$$

Here $\tilde{\beta} \in \mathbb{R}^p$ is an additional estimate of β_0 that may be computed using the auxiliary dataset, or in certain circumstances may be taken as $\hat{\beta}$ itself: we discuss these two cases in the following sections. Given the vector R of residuals, the basic form of our test statistic is w^TR; here $w \in \mathbb{R}^n$ is a direction typically derived using the auxiliary dataset. We describe in detail the construction of such a w in Section 2.1, where the goal is general goodness-of-fit testing.

A further modification of the method can allow us to use multiple directions w to test simultaneously for different departures from the null or to aggregate over different directions derived using flexible regression methods with different tuning parameters. Given a set $W \subseteq \mathbb{R}^n$ of direction vectors w, our proposed test statistic then takes the form

$$\sup_{w\in W} w^TR.$$

We illustrate the use of this more general form of our test statistic for testing the significance of a group of variables. Such a problem may not immediately seem like goodness-of-fit testing, but is equivalent to testing the adequacy of a model not involving the group of variables in question. We explain how this may be addressed by our framework in Section 2.2 and describe a wild bootstrap procedure (Wu, 1986; Chernozhukov et al., 2013) to approximate the distribution of the test statistic under the null.

2.1 Goodness-of-fit testing

To motivate our general procedure for goodness-of-fit testing, consider the vector R_{ora} of Pearson residuals with an oracle variance scaling, whose i-th component is given by

$$R_{\text{ora},i} := \frac{Y_i - \mu(x_i^T \hat{\beta})}{D_{\beta_0,ii}}, \quad i = 1, \ldots, n,$$

where $D^2_{\beta_0,ii} := V(\mu(x_i^T \beta_0))$. We may decompose the residuals into noise and estimation error terms by writing

$$R_{\text{ora},i} = \varepsilon_i + r_i, \tag{1}$$

where $\varepsilon_i := \{Y_i - \mu(x_i^T \beta_0)\}/D_{\beta_0,ii}$ and $r_i := \{\mu(x_i^T \beta_0) - \mu(x_i^T \hat{\beta})\}/D_{\beta_0,ii}$. If the generalized linear model is correct, then $\mathbb{E}(\varepsilon_i | x_i) = 0$ and $\text{Var}(\varepsilon_i | x_i) = 1$. Turning to the remainder term r_i, a first-order Taylor expansion of μ yields the approximation

$$r_i \approx \frac{\mu'(x_i^T \beta_0)}{D_{\beta_0,ii}} x_i^T (\beta_0 - \hat{\beta}).$$

Writing Ω_0 for the diagonal matrix with entries $\mu'(x_i^T \beta_0)/D_{\beta_0,ii}$ for $i = 1, \ldots, n$, and $\varepsilon := (\varepsilon_1, \ldots, \varepsilon_n)$, we obtain the decomposition

$$R_{\text{ora}} \approx \varepsilon + \Omega_0 X (\beta_0 - \hat{\beta}). \tag{2}$$

Consider a unit vector $w \in \mathbb{R}^n$; as discussed in Section 1.1, this will typically be constructed from an application of a residual prediction method on the auxiliary data. Our oracle $w^T R_{\text{ora}}$ then satisfies

$$w^T R_{\text{ora}} \approx w^T \varepsilon + w^T \Omega_0 X (\beta_0 - \hat{\beta}) =: w^T \varepsilon + \delta. \tag{3}$$

Under suitable conditions on w and on the moments of the errors, the Berry–Esseen theorem should ensure that the pivot term $w^T \varepsilon$ is well approximated by a standard Gaussian random variable. To keep the remainder term δ in (3) under control we can leverage the fact that under the null, we can expect $\|\hat{\beta} - \beta_0\|_1$ to be small. If w satisfies a near-orthogonality condition

$$\|X^T \Omega_0 w\|_\infty \leq C \sqrt{\log p}, \tag{4}$$

for some $C > 0$, then Hölder's inequality will yield $|\delta| \leq C \sqrt{\log p} \|\hat{\beta} - \beta_0\|_1$, which asymptotically vanishes under suitable conditions on the sparsity of β_0.

To guarantee the near-orthogonality condition (4), we may use the square-root Lasso (Belloni et al., 2011; Sun and Zhang, 2012): for $\lambda_{\text{sq}} > 0$, let

$$\hat{\beta}_{\text{ora-sq}} := \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{\sqrt{n}} \|\Omega_0(\hat{f}(X) - X\beta)\|_2 + \lambda_{\text{sq}} \|\beta\|_1 \right\}.$$

The Karush–Kuhn–Tucker (KKT) conditions for the convex programme imply that the resulting vector of scaled residuals,

$$w_{\text{ora}} := \frac{\Omega_0(\hat{f}(X) - X\hat{\beta}_{\text{ora-sq}})}{\|\Omega_0(\hat{f}(X) - X\hat{\beta}_{\text{ora}-\text{sq}})\|_2},$$

satisfies the near-orthogonality property $\|X^T \Omega_0 w_{\text{ora}}\|_\infty \leq C\sqrt{\log p}$ when $\lambda_{\text{sq}} = C\sqrt{\log p/n}$. Note that in performing this square-root Lasso regression, we are not assuming that $\hat{f}$ is well-approximated by a sparse linear combination of variables: we are simply exploiting the stationarity properties of the solution to the square-root Lasso optimisation problem[1].

From the reasoning above, we conclude that, under appropriate conditions, a simple test based on the asymptotic normality of $w_{\text{ora}}^T R_{\text{ora}}$ will keep the type I error under control. In order to create a version of the test statistic that does not require oracular knowledge of Ω_0, we may replace this quantity with variance estimates based either on $\hat{\beta}$ or on an estimate derived from the auxiliary data; we use the latter approach as this simplifies the analysis. The overall procedure is summarised in Algorithm 1 below.

Algorithm 1. Goodness-of-fit testing.

Input: sample $(X, Y) \in \mathbb{R}^{n \times p} \times \mathcal{Y}^n$; auxiliary sample $(X_A, Y_A) \in \mathbb{R}^{n \times p} \times \mathcal{Y}^n$; $\lambda, \lambda_A, \lambda_{\text{sq}} > 0$.

1: **Estimation:** Fit a GLM Lasso to (X, Y) and (X_A, Y_A) (with tuning parameters λ, λ_A respectively) yielding estimators $\hat{\beta}$ and $\hat{\beta}_A$, respectively.

2: **Residual prediction:** Compute the residuals $Y_A - \mu(X_A\hat{\beta}_A)$ and fit a flexible regression method of these residuals versus X_A to obtain a prediction function $\hat{f} : \mathbb{R}^{n \times p} \to \mathbb{R}^n$.

3: **Near orthogonalization:** Let $\hat{D}_A := \operatorname{diag}\big(V^{1/2}(\mu(x_i^T \hat{\beta}_A))_{i=1}^n\big)$, construct the diagonal weight matrix $\hat{\Omega}_A := \operatorname{diag}\big(\big(\frac{\mu'(x_i^T \hat{\beta}_A)}{\hat{D}_{A,ii}}\big)_{i=1}^n\big)$ and compute an approximate projection of the prediction $\hat{\Omega}_A \hat{f}(X)$ onto the column space of $\hat{\Omega}_A X$:

$$\hat{\beta}_{\text{sq}} := \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{\sqrt{n}} \|\hat{\Omega}_A(\hat{f}(X) - X\beta)\|_2 + \lambda_{\text{sq}} \|\beta\|_1 \right\}. \tag{5}$$

Define a direction

$$\hat{w}_A := \frac{\hat{\Omega}_A(\hat{f}(X) - X\hat{\beta}_{\text{sq}})}{\|\hat{\Omega}_A(\hat{f}(X) - X\hat{\beta}_{\text{sq}})\|_2}. \tag{6}$$

4: **Test statistic:** Compute the residual vector $\hat{R} := \hat{D}_A^{-1}(Y - \mu(X\hat{\beta}))$ and let $T := \hat{w}_A^T \hat{R}$.

---

[1] In principle, there is a possibility that we obtain a degenerate solution with $\hat{f}(X) = X\hat{\beta}_{\text{ora-sq}}$. However, we can observe directly whether or not this occurs, and have never seen this happen in any of our numerical experiments.

Output: $p_{\text{value}} = 1 - \Phi(T)$

In practice, the auxiliary dataset (X_A, Y_A) would be obtained through sample splitting. The effect of the randomness induced by such a split can be mitigated using methods designed to aggregate over multiple sample splits, as studied for instance in Meinshausen et al. (2009).

Remark 1. In order to achieve better control of the type I error, in practice it may be helpful to use a slight modification of the procedure proposed in Algorithm 1. Define $\hat{S} := \{j : \hat{\beta}_j \neq 0\}$. Rather than calculating the direction $\hat{w}_A$ through $\hat{\beta}_{\text{sq}}$, we may instead use

$$\tilde{\beta}_{\text{sq}} := \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \left\{ \frac{1}{\sqrt{n}} \left\| \hat{\Omega}_A \left(\hat{f}(X) - X\beta \right) \right\|_2 + \lambda_{\text{sq}} \|\beta_{\hat{S}^c}\|_1 \right\}$$

in its place within (6). In this way, we enforce that $\hat{w}_A$ is *exactly* orthogonal to $\hat{\Omega}_A X_{\hat{S}}$. This helps to keep the remainder term arising from the asymptotic expansion of the test statistic under control, as can be seen from the following argument. Assume that a "beta-min condition" is satisfied, that is for all $j \in S := \{k \in \{1, \ldots, p\} : \beta_{0,k} \neq 0\}$ it holds that $\beta_{0,j} \gtrsim \sqrt{s(\log p)/n}$, where $s := |S|$. Then, under mild additional conditions, $\hat{S} \supseteq S$ with high probability (e.g. Bühlmann and van de Geer, 2011, Corollary 7.6). On the event that this occurs, we have that the remainder term in (1) satisfies

$$\begin{aligned} \hat{w}_A^T \Omega_0 X (\beta_0 - \hat{\beta}) &\approx \hat{w}_A^T \hat{\Omega}_A X (\beta_0 - \hat{\beta}) \\ &= \hat{w}_A^T \hat{\Omega}_A X_{\hat{S}} (\beta_{0,\hat{S}} - \hat{\beta}_{\hat{S}}) = 0. \end{aligned}$$

Even without such a beta-min condition, it is plausible that we will obtain a reduction in this bias term through this strategy of exact orthogonalization. In fact, this form of exact orthogonalization may be useful for other procedures, such as the debiased Lasso.

2.2 Group testing

Our framework of residual prediction tests also encompasses significance testing of groups of regression coefficients in a generalized linear model. In this section, it is convenient to work with $\mu(\cdot)$ being the canonical link function. Suppose that we wish to test $H_0 : \beta_G = 0$ for a given group $G \subseteq \{1, \ldots, p\}$. We first form the vector $\hat{R}_G$ of residuals based on a GLM Lasso fit of Y on X_{-G}. Then, rather than constructing a single direction w using an auxiliary dataset, we can use multiple directions given by the columns of X_G. Specifically, we use the test statistic $\max_{j \in G} |\hat{w}_j^T \hat{R}_G|$ where $\hat{w}_j$ is given by the scaled residuals of the weighted square-root Lasso regressions of X_j on to X_{-G}.

Note that under the null, X_G will be independent of the noise ε (1), and so sample splitting is not necessary in this case to mitigate the potentially complicated dependence of the directions and residuals $\hat{R}_G$. The limiting distribution of the test however will not be Gaussian due to the maximisation over multiple directions. Instead, we argue that $\max_{j \in G} |\hat{w}_j^T \hat{R}_G| \approx \max_{j \in G} |\hat{w}_j^T \varepsilon|$ and then use a wild bootstrap procedure to approximate the distribution of this latter quantity. The overall procedure is summarised in Algorithm 2 below.

Algorithm 2. Group Test.

Input: Group $G \subseteq \{1, \dots, p\}$; sample $(X, Y) \in \mathbb{R}^{n \times p} \times \mathbb{R}^n$; $B \in \mathbb{N}$; $\lambda, \lambda_{\text{nw}} > 0$.

1: Fit a GLM Lasso to (X_{-G}, Y) with a tuning parameter λ to obtain an estimator $\hat{\beta}_{-G} \in \mathbb{R}^{p-|G|}$. Let $\hat{D}^2 := \text{diag}(\mu'(X\hat{\beta}_{-G}))$. Compute the vector of residuals $\hat{R}_G := \hat{D}^{-1}(Y - \mu(X_{-G}\hat{\beta}_{-G}))$.

2: For each $j \in G$, compute the nodewise regression estimator

$$\hat{\gamma}_j := \underset{\gamma \in \mathbb{R}^{p-|G|}}{\text{argmin}} \frac{1}{\sqrt{n}} \|\hat{D}(X_j - X_{-G}\gamma)\|_2 + \lambda_{\text{nw}} \|\gamma\|_1,$$

and let

$$\hat{w}_j := \frac{\hat{D}(X_j - X_{-G}\hat{\gamma}_j)}{\|\hat{D}(X_j - X_{-G}\hat{\gamma}_j)\|_2}.$$

3: For $j \in G$, let $T_j := \hat{w}_j^T \hat{R}_G$ and evaluate the test statistic $T := \max_{j \in G} |T_j|$.

4: For $b = 1, \dots, B$ generate independent random variables $e_1^b, \dots, e_n^b \sim \mathcal{N}(0, 1)$ and let

$$T^b := \max_{j \in G} \left| \sum_{i=1}^n \hat{w}_{j,i} \hat{R}_{G,i} e_i^b \right|,$$

where $\hat{w}_{j,i}$ and $\hat{R}_{G,i}$ are the i-th entries of $\hat{w}_j$ and $\hat{R}_G$, respectively.

5: Calculate the p-value

$$p_{\text{value}} := \frac{1}{B+1} \left(1 + \sum_{b=1}^B \mathbb{1}_{\{T^b \geq T\}} \right).$$

Output: p_{value}

Our Algorithm 2 is similar to the de-biased Lasso for generalized linear models (van de Geer et al., 2014). The main difference however is that the de-biased Lasso aims to ensure the directions $\hat{w}_j$ are almost orthogonal to X_{-j}, whereas we only impose near-orthogonality with respect to X_{-G}. Thus for large groups G, more of the direction of X_G is preserved in the $\hat{w}_j$, which typically leads to better power of the test.

Remark 2. In practice, one needs to specify the tuning parameters $\lambda, \lambda_A, \lambda_{\text{sq}}$ and λ_{nw} in Algorithms 1 and 2. A popular technique is 10-fold cross-validation, which we also use to select λ and λ_A for the GLM Lasso in our empirical experiments. For the square-root Lasso optimization problem (λ_{sq} and λ_{nw}), we adopted the approach from Sun and Zhang (2012) that proposes a universal choice $\lambda_{\text{sq}} = \lambda_{\text{nw}} = \sqrt{2(\log p)/n}$. We provide a more detailed discussion on how to obtain theoretical guarantees for these choices in Section 3.

3 Theoretical guarantees

In this section we provide theoretical guarantees for the tests proposed in Algorithms 1 and 2. Throughout, we assume that $Y_1, \dots, Y_n$ are conditionally independent given X, with the conditional distribution of each Y_i only depending on X through x_i (in fact, in Section 3.1.2, we will make the stronger assumption that $(Y_i, x_i)_{i=1}^n$ are independent and identically distributed).

3.1 Size of the test

In the following, we show that under the null hypothesis, the size of the test is asymptotically correct. We explore goodness-of-fit testing in Section 3.1.1 and group testing in Section 3.1.2.

3.1.1 Goodness-of-fit testing

Here we show that our test statistic has a Gaussian limiting distribution and we establish a bound on the type I error of the test. Our result makes use of the fact that when the model is well specified, the GLM Lasso performs well in terms of estimation. Specifically, under certain conditions, it holds with high probability that $\hat{\beta} \in \Theta(\lambda, \beta_0, X)$, where $\Theta(\lambda, \beta_0, X)$ is a local neighbourhood of β_0 defined by

$$\Theta(\lambda, \beta_0, X) := \left\{\vartheta \in \mathbb{R}^p : \|\vartheta - \beta_0\|_1 \leq s\lambda, \|X(\vartheta - \beta_0)\|_2^2/n \leq s\lambda^2\right\},$$

with $s := \|\beta_0\|_0$ as the number of non-zero entries of β_0; see for example Bühlmann and van de Geer (2011, Corollary 6.3). Sufficient conditions for this to occur include $\lambda \asymp \sqrt{\log p/n}$, $s = o(n/\log p)$ and further conditions on the tail behaviour of the errors $Y_i - \mu(x_i^T\beta_0)$, the design matrix X and the link function, as detailed below.

Condition 1. *Assume that* $\mathbb{E}(Y_i|x_i = x) = \mu(x^T\beta_0)$ *and that* $\mathrm{Var}(Y_i|x_i = x) = V(\mu(x^T\beta_0))$, *that the inverse link function* $u \mapsto \mu(u)$ *is differentiable and* $\mu'(u) > 0$ *for all* $u \in \mathbb{R}$. *Assume that* $\max_{i=1,\ldots,n} |x_i^T\beta_0| \leq K_0/2$ *and hat the maps* $u \mapsto \mu'(u)$ *and* $u \mapsto V(\mu(u))$ *are* L*-Lipschitz on the interval* $[-K_0, K_0]$. *Suppose moreover that the weights satisfy* $\min_i D_{\beta_0,ii} \geq d_{\min}$ *for some constant* $d_{\min} > 0$. *Assume that* $\mathbb{E}\{|Y_i - \mu(x_i^T\beta_0)|^3/D_{\beta_0,ii}^3 \mid X\} \leq C_\varepsilon$ *for some constant* $C_\varepsilon > 0$, *that* $\max_{i=1,\ldots,n} \|x_i\|_\infty \leq K_X$, *that* $12d_{\min}^{-2}LK_Xs\lambda_A \leq 1$ *and that* $K_Xs\max(\lambda, \lambda_A) \leq K_0/2$.

Condition 1 is satisfied for generalized linear models with canonical links under mild additional conditions. For example, in the case of logistic regression, the condition on the weights is satisfied if the class probability $\pi_0(x) = \mathbb{P}(Y_i = 1|x_i = x)$ is bounded away from zero and one. Boundedness of the design (along with other conditions, including $12d_{\min}^{-2}LK_Xs\lambda_A \leq 1$) guarantees that the weights can be consistently estimated. For our result below, it is convenient to introduce the shorthand notation $Z_A := (X, X_A, Y_A)$.

Theorem 1. *Consider Algorithm 1 with tuning parameters* $\lambda, \lambda_A, \lambda_{\mathrm{sq}} > 0$. *Assume that Condition 1 is satisfied, that* $\hat{\beta}_A \in \Theta(\lambda_A, \beta_0, X_A)$ *and let*

$$\delta := \mathbb{P}\big(\hat{\beta} \notin \Theta(\lambda, \beta_0, X) \mid X\big).$$

Then there exists a constant[2] $C > 0$ *such that whenever* Z_A *satisfies* $\hat{f}(X) \neq X\hat{\beta}_{\mathrm{sq}}$, *we have for any* $z \in \mathbb{R}$ *that*

$$|\mathbb{P}(T \leq z|Z_A) - \Phi(z)| \leq \delta + C\big\{\lambda_{\mathrm{sq}}\sqrt{n}s\lambda + L\|\hat{w}_A\|_\infty s(\lambda^2 + \lambda_A^2)n + LK_Xs\lambda_A + \|\hat{w}_A\|_\infty\big\}, \quad (7)$$

where Φ *denotes the standard normal distribution function.*

[2]Here and below, the constants in the conclusions of our results may depend upon quantities introduced as constants in the relevant conditions for these results.

We now discuss the terms on the right-hand side of (7). The terms $\lambda_{\text{sq}}\sqrt{n}s\lambda$ and $LK_X s\lambda_A$ arise from bounding the bias term (near-orthogonalization step) and from bounding the weights, respectively. The presence of the $\|\hat{w}_A\|_\infty$ term in the bound stems from the contribution of each individual component $\hat{w}_{A,i}$ to the variance of the pivot term and that of the higher-order terms omitted in (2), which create a bias in the distribution of the test statistic.

To provide some intuition on the size of $\|\hat{w}_A\|_\infty$, recall that $\hat{w}_A$ is a vector in $\mathbb{R}^n$ with $\|\hat{w}_A\|_2 = 1$, so we may hope for $\|\hat{w}_A\|_\infty$ to be small; in fact, it can be shown in certain settings, and under additional technical conditions, that $\|\hat{w}_A\|_\infty = \mathcal{O}_P(\log n/\sqrt{n})$. In that case, for the asymptotically optimal choice of tuning parameters $\lambda \asymp \lambda_A \asymp K_X\sqrt{\log p/n}$ and $\lambda_{\text{sq}} \asymp \sqrt{\log p/n}$, the bound in Theorem 1 when $K_X \geq 1$, $p \geq 2$ and $s \geq 1$ reduces to

$$|\mathbb{P}\left(T \leq z|Z_A\right) - \Phi(z)| = \mathcal{O}_P\left(\delta + \frac{sLK_X^2(\log n)\log p}{\sqrt{n}}\right). \tag{8}$$

We also remark that we observe $\hat{w}_A$, and if the size of its ℓ_∞-norm is a concern, then we can modify the square-root Lasso objective to control it explicitly. Indeed, consider setting

$$(\tilde{\beta}_{\text{sq}}, \tilde{\eta}_{\text{sq}}) := \underset{(\beta,\eta)\in\mathbb{R}^p\times\mathbb{R}^n}{\operatorname{argmin}} \left\{\frac{1}{\sqrt{n}}\|\hat{\Omega}_A(\hat{f}(X) - X\beta) - \eta\|_2 + \lambda_{\text{sq}}\|\beta\|_1 + \lambda_\eta\|\eta\|_1\right\}$$

and let

$$\tilde{w}_A := \frac{\hat{\Omega}_A(\hat{f}(X) - X\tilde{\beta}_{\text{sq}}) - \tilde{\eta}_{\text{sq}}}{\|\hat{\Omega}_A(\hat{f}(X) - X\tilde{\beta}_{\text{sq}}) - \tilde{\eta}_{\text{sq}}\|_2}.$$

Then the KKT conditions of the optimisation problem imply in particular both that a near-orthogonality condition similar to (4) is satisfied for suitable λ_{sq}, and also that $\|\tilde{w}_A\|_\infty \leq \sqrt{n}\lambda_\eta$. Our empirical results in Section 4 however suggest that in practice $\|\hat{w}_A\|_\infty$ typically satisfies the necessary constraint and therefore we propose to use the simpler standard square-root Lasso without the above modifications.

Remark 3. A natural question concerns the extent to which the practical choice of the tuning parameters affects our results and whether it guarantees that we achieve the desired rates of convergence as in (8). We make use of cross-validation, and recent work (Chetverikov and Chernozhukov, 2016) has shown that cross-validation with the Lasso indeed leads to nearly optimal (up to a logarithmic factor) rates of convergence for the squared prediction error and the ℓ_1-norm of the estimation error. We note that the work of Chetverikov and Chernozhukov (2016) covers only the linear model, while our paper covers a more general setting. Understanding the performance of cross-validated procedures in more general models is an important avenue for future research, but is beyond the scope of this paper.

3.1.2 Group testing

In this section, we derive theoretical properties for the group testing procedure proposed in Algorithm 2. Since we do not use sample splitting, we cannot directly apply the arguments of Theorem 1, as the direction $\hat{w}_j$ depends on $\hat{\beta}_{-G}$ via the weights $\hat{D} = D_{\hat{\beta}_{-G}}$. In order to understand this dependence, here we consider the setting of random bounded design and assume the response–covariate pairs (Y_i, x_i) are all independent and identically distributed.

We aim to use the multiplier bootstrap procedure (Chernozhukov et al., 2013) to estimate the distribution of the test statistic $\max_{j\in G}|T_j|$ as described in Algorithm 2, but we first summarize a preliminary result which shows that, under appropriate conditions, T_j can be asymptotically approximated by the zero-mean average $w_j^T\varepsilon$. Here we define $w_j := D_{\beta_0}(X_j - X_{-G}\gamma_{0,j})/(\sqrt{n}\tau_j)$, where

$$\tau_j^2 := \frac{1}{n}\mathbb{E}\|D_{\beta_0}(X_j - X_{-G}\gamma_{0,j})\|_2^2,$$

and $\gamma_{0,j}$ is the population version of $\hat{\gamma}_j$ from Algorithm 2; i.e.

$$\gamma_{0,j} := \operatorname*{argmin}_{\gamma\in\mathbb{R}^{p-|G|}} \frac{1}{n}\mathbb{E}\|D_{\beta_0}(X_j - X_{-G}\gamma)\|_2^2.$$

Recall that

$$\varepsilon = D_{\beta_0}^{-1}\big(Y - \mu(X\beta_0)\big).$$

In order to guarantee consistency of $\hat{\gamma}_j$ in Algorithm 2, we will introduce the additional requirement that $\gamma_{0,j}$ is sparse. Denote by $\beta_{0,-G} \in \mathbb{R}^{p-|G|}$ the subset of components of β_0 corresponding to indices in G^c. We also define

$$\Theta_{-G}(\lambda,\beta_0) := \left\{\vartheta\in\mathbb{R}^{p-|G|} : \|\vartheta - \beta_{0,-G}\|_1 \le s\lambda, \|X_{-G}(\vartheta-\beta_{0,-G})\|_2^2/n \le s\lambda^2\right\}.$$

Condition 2.

(i) *Let* $\eta_i := Y_i - \mu(x_i^T\beta_0)$ *and assume that there exist constants* $c_1, c_2, c_3 > 0$ *such that*

$$\mathbb{E}(e^{\eta_i^2/c_1^2}|X) \le c_2 \quad \textit{and} \quad \mathbb{E}(\eta_i^2|X) \ge c_3,$$

for all $i = 1,\dots,n$.

(ii) *There exists* $K \ge 1$ *such that* $\|X\|_\infty \le K$ *and* $\max_{j\in G}\|X_{-G}\gamma_{0,j}\|_\infty \le K$.

(iii) *For some* $\delta > 0$ *and all* $\beta\in\mathbb{R}^p$ *satisfying* $\|\beta-\beta_0\|_1 \le \delta$ *it holds that* $c_0 \le \mu'(x^T\beta) \le C_0$ *for some constants* $c_0, C_0 > 0$ *and all* $x\in\mathbb{R}^p$ *with* $\|x\|_\infty \le K$.

(iv) *Denoting* $\Sigma_0 := \mathbb{E}X^T D_{\beta_0}^2 X/n$, *we have* $1/\Lambda_{\min}(\Sigma_0) \le C_e$ *and* $\|\Sigma_0\|_\infty \le C_e$ *for some constant* $C_e > 0$.

(v) *We have* $\max_{j\in G}\|\gamma_{0,j}\|_0 \le s$, $\|\beta_0\|_0 \le s$ *and there exists a sequence* (a_n) *with* $a_n \to 0$ *and* $K^3 s\log p/\sqrt{n} \le a_n$.

Proposition 1. *Assume that Conditions 1 and 2 are satisfied with* $\mu(\cdot)$ *being the canonical link and assume that* $\hat{\beta}_{-G}$ *satisfies*

$$\mathbb{P}\big(\hat{\beta}_{-G}\in\Theta_{-G}(\lambda,\beta_0)\big) \le 1/p. \tag{9}$$

Consider $T_j, j\in G$ *as defined in Algorithm 2 with tuning parameters* $\lambda \asymp \sqrt{\log p/n}$ *and* $\lambda_{\text{nw}} \asymp K\sqrt{\log p/n}$. *Assume that the null hypothesis* $H_0 : \beta_{0,G} = 0$ *holds. Then there exists a constant* $C > 0$ *such that with probability at least* $1 - 3/p$, *we have*

$$\max_{j\in G}|T_j - w_j^T\varepsilon| \le CK^3\frac{s\log p}{\sqrt{n}}.$$

Using Proposition 1 and the results of Chernozhukov et al. (2013), we can show that the quantiles of $\max_{j\in G}|T_j|$ can be approximated by the quantiles of $T^b := \max_{j\in G}|\sum_{i=1}^n \hat{w}_{j,i}\hat{R}_{G,i}e_i^b|$ where $e_1^b,\dots,e_n^b$ are independent $\mathcal{N}(0,1)$ random variables. We only need to guarantee that T_j is well-approximated by $w_j^T\varepsilon$ and we pay a price of $\log|G|$ for testing $|G|$ hypothesis simultaneously, where $|G|$ denotes the cardinality of G.

Define the α-quantile of T^b conditional on $(x_i,Y_i)_{i=1}^n$ by

$$c_{T^b}(\alpha) := \inf\{t\in\mathbb{R} : \mathbb{P}_e(T^b\le t)\ge\alpha\},$$

where $\mathbb{P}_e$ is the probability measure induced by the multiplier variables $(e_i^b)_{i=1}^n$ holding $(x_i,Y_i)_{i=1}^n$ fixed.

Theorem 2. *Assume the conditions of Proposition 1 and that there exist constants $C_2,c_2>0$ such that*

$$\max\left\{K^3\frac{s\log p}{\sqrt{n}}\log(2|G|)+4/p,\ \ K^4\log(2|G|n)^7/n\right\}\le C_2n^{-c_2}. \tag{10}$$

Then there exist constants $c,C>0$ such that

$$\sup_{\alpha\in(0,1)}\left|\mathbb{P}\left(\max_{j\in G}|T_j|<c_{T_b}(\alpha)\right)-\alpha\right|\le Cn^{-c}.$$

Theorem 2 shows that if the generalized linear model is correct and the null hypothesis $\beta_{0,G}=0$ holds, then Algorithm 2 produces an asymptotically valid p-value for testing this hypothesis.

3.2 Power analysis for goodness-of-fit testing

The choice of w as postulated in Theorem 1 guarantees that the type I error for goodness-of-fit testing stays under control. We now provide guarantees on the local power of the test. To this end, let us suppose that the true model has conditional expectation function $m_0(x)=\mu\big(x^T\beta_0+g_0(x)\big)$. Here g_0 represents a small nonlinear perturbation of the linear predictor $x^T\beta_0$. Our aim here is to understand how this propogates through to the distribution of our test statistic. We will suppose that the perturbation is small enough that GLM Lasso estimates lie with high probability within local neighbourhoods of β_0. Let us first provide some intuition on the expected value of the test statistic under model misspecification. Writing $f_0(x)=x^T\beta_0+g_0(x)$, the expectation of the theoretical residuals $\varepsilon=D_{\beta_0}^{-1}\big(Y-\mu(X\beta_0)\big)$ is given by

$$w_0 := \mathbb{E}\varepsilon = \mathbb{E}D_{\beta_0}^{-1}\big(\mu(f_0(X))-\mu(X\beta_0)\big).$$

As argued in Section 2, the oracular test statistic w^TR_{ora} can be approximated by the scalar product $w^T\varepsilon$. Therefore, in order to obtain good power properties, we should seek to construct a direction w so as to maximize $\mathbb{E}w^T\varepsilon$. The oracular choice $w:=w_0/\|w_0\|_2$ yields by a Taylor expansion the approximation

$$\mathbb{E}w^T\varepsilon=\|w_0\|_2\approx\|D_{\beta_0}(f_0(X)-X\beta_0)\|_2=\|D_{\beta_0}g_0(X)\|_2.$$

We therefore see that the test statistic behaves in expectation as a weighted ℓ_2-norm of the nonlinear term $g_0(X)$. We now provide a theoretical justification which can be used for local asymptotic guarantees on the power of our method. We introduce the following conditions which are modifications of Condition 1 to account for the case when the model is misspecified.

Condition 3. *Assume that* $\mathbb{E}(Y_i|x_i = x) = \mu(f_0(x))$*, that the inverse link function* $u \mapsto \mu(u)$ *is differentiable and* $\mu'(u) > 0$ *for all* $u \in \mathbb{R}$*. Assume that* $\max_{i=1,\dots,n} |x_i^T \beta_0| \le K_0/2$ *and that the maps* $u \mapsto \mu'(u)$ *and* $u \mapsto V(\mu(u))$ *are Lipschitz on the interval* $[-K_0, K_0]$ *with a parameter* L*. Suppose moreover that the weights satisfy* $\min_i D_{\beta_0,ii} \ge d_{\min}$ *for some constant* $d_{\min} > 0$*. Assume that* $\mathbb{E}\big(|Y_i - \mu(f_0(x_i))|^3/D^3_{Y,ii} \;\big|\; X\big) \le C_\varepsilon$ *for a constant* $C_\varepsilon > 0$*, where we denote* $D^2_Y := \mathrm{Cov}(Y|X)$*. Let* $\max_{i=1,\dots,n} \|x_i\|_\infty \le K_X$*, assume that* $12 d_{\min}^{-2} L K_X s\lambda \le 1$*, that* $|D^2_{Y,ii} D^{-2}_{\beta_0,ii} - 1| \le 2 d_{\min}^{-2} L K_X s\lambda$ *and that* $K_X s \max(\lambda, \lambda_A) \le K_0/2$*.*

Theorem 3. *Consider Algorithm 1 with tuning parameters* $\lambda, \lambda_A, \lambda_{\mathrm{sq}}$*. Assume Condition 3, that* $\hat\beta_A \in \Theta(\lambda_A, \beta_0, X_A)$ *and let*

$$\delta := \mathbb{P}(\hat\beta \notin \Theta(\lambda, \beta_0, X)|X). \tag{11}$$

Then there exists a constant $C > 0$ *such that, whenever* Z_A *is such that* $\hat f(X) \ne X\hat\beta_{\mathrm{sq}}$*, we have for any* $z \in \mathbb{R}$ *that*

$$\big|\mathbb{P}\left(T - \boldsymbol{\Delta} < z|Z_A\right) - \Phi(z)\big| \le \delta + C\big\{\lambda_{\mathrm{sq}}\sqrt{n} s\lambda + \|\hat w_A\|_\infty s(\lambda^2 + \lambda_A^2)n + K_X s\lambda_A + \|\hat w_A\|_\infty\big\}, \tag{12}$$

where

$$\boldsymbol{\Delta} := \hat w_A^T \hat D_A^{-1}\big\{\mu\big(f_0(X)\big) - \mu(X\beta_0)\big\}.$$

Under the null hypothesis, we have $\boldsymbol{\Delta} = 0$ and $D_Y = D_{\beta_0}$; thus we recover the result of Theorem 1. The departure of the model from the null hypothesis is captured by $\boldsymbol{\Delta}$. Hence the theorem shows that to detect departures from the null, the direction w must be "correlated" with the signal that remains in the residuals under misspecification, namely $\mu(f_0(X)) - \mu(X\beta_0)$. Under a local alternative, e.g. $f_0(X) = X\beta_0 + g_0(X)/\sqrt{n}$ where $\|g_0(X)\|_2 = 1$, we have $\boldsymbol{\Delta} \asymp 1$ provided the angle between $\hat w_A$ and the remaining signal is bounded away from zero.

Theorem 3 relies on rates of convergence of the "projected" estimator $\hat\beta$ in (11) when the model is misspecified. Oracle inequalities for Lasso-regularized estimators in high-dimensional settings have been well explored; we refer to Bühlmann and van de Geer (2011) and the references therein. If there is misspecification, we hope that the projected estimator behaves as if it knows which variables are relevant for a linear approximation of the possibly nonlinear target f_0. In Appendix A.4, we summarize how misspecification affects estimation of the best linear approximation, based on the approach of Bühlmann and van de Geer (2011). These results guarantee that under a local alternative, the Lasso for generalized linear models still satisfies the condition

$$\mathbb{P}\big(\hat\beta \in \Theta(\lambda, \beta_0, X)|X\big) \to 1.$$

3.3 Consequences for logistic regression

In this section we show how our general theory applies to the problem of goodness-of-fit testing for logistic regression models. We take $\mathcal{Y} = \{0, 1\}$ and assume $(Y_i|x_i = x) \sim \mathrm{Bernoulli}(\pi_0(x))$. Define

$$f_0(x) := \log\left(\frac{\pi_0(x)}{1 - \pi_0(x)}\right),$$

that is, $\mathbb{E}(Y_i|x_i = x) = \pi_0(x) = \mu(f_0(x))$, for the inverse link function $\mu(u) = 1/(1 + e^{-u})$. The function f_0 may be potentially nonlinear in x. The ℓ_1-regularized logistic regression estimator is

$$\hat{\beta} := \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{-Y_i x_i^T \beta + d(x_i^T \beta) + \lambda \|\beta\|_1\},$$

where $d(\xi) := \log(1 + e^{\xi})$. Let $\beta_0 \in \mathbb{R}^p$ be the best approximation obtained by a GLM (Bühlmann and van de Geer, 2011, Section 6.3, p. 115). We define $S := \{j : \beta_{0,j} \neq 0\}$ and $s := |S|$.

Corollary 1 below follows by combining Theorem 3 with existing results on ℓ_1-penalized logistic regression. We assume conditions, stated formally in Lemma 6 in Appendix A, which guarantee that with high probability this penalized estimator is sufficiently close to β_0.

Corollary 1. *Assume that the conditions of Lemma 6 in Appendix A hold and in addition assume that* $\hat{\beta}_A \in \Theta(\lambda, \beta_0, X_A)$ *and that* $12 c_0^{-2} K s \lambda_A \leq 1$ *where* $c_0^2 = (e^{\eta}/\epsilon_0 + 1)^{-2}$ *and* K, η *and* ϵ_0 *are defined in Lemma 6. Suppose that* $\lambda \asymp \lambda_{\text{sq}} \asymp \lambda_A \asymp \sqrt{\log(2p)/n}$. *Then there exists a constant* $C > 0$ *such that for any* $z \in \mathbb{R}$, *and whenever* Z_A *is such that* $\hat{f}(X) \neq X\hat{\beta}_{\text{sq}}$,

$$\left| \mathbb{P}\left(T - \boldsymbol{\Delta} < z | Z_A\right) - \Phi(z) \right| \leq C \left((2p)^{-1} + \frac{s\{\log(2p) + K\sqrt{\log(2p)}\}}{\sqrt{n}} + \|\hat{w}_A\|_\infty s \log(2p) \right), \quad (13)$$

where

$$\boldsymbol{\Delta} := \hat{w}_A^T \hat{D}_A^{-1} \{\mu\big(f_0(X)\big) - \mu(X\beta_0)\}.$$

4 Empirical results: Logistic regression

In this section we explore the empirical performance of the methods for goodness-of-fit testing and group testing in the setting of logistic regression. We begin by considering goodness-of-fit testing in low-dimensional settings in Section 4.1 and in high-dimensional settings in Section 4.2. Goodness-of-fit testing on semi-real data is investigated in Section 4.3, while in Section 4.4, we explore group testing in high-dimensional settings. The code to reproduce the results in this section is available at `https://github.com/janajankova/GRPtests` and in the Online Supplementary Material.

4.1 Low-dimensional settings

While for low-dimensional settings, there are numerous methods available for testing goodness-of-fit as discussed in Section 1.2, we show that our test from Algorithm 1 may be advantageous even here. We compare the performance of our test (with residual prediction method being a random forest with default tuning parameter choices) against the Hosmer–Lemeshow $\hat{C}$ test, the Hosmer–Lemeshow $\hat{H}$ test (see Lemeshow and Hosmer Jr (1982)) and the le Cessie–van Houwelingen–Copas–Hosmer unweighted sum of squares test (see Hosmer et al. (1997)). These tests are implemented in the function `HLgof.test()` in the R package `MKmisc` (Kohl, 2018).

We simulated data from a logistic regression model with sample size $N = 300$ and $p = 10$ covariates according to

$$Y_i | x_i = u \sim \text{Bern}\big(\pi(u)\big),$$

where

$$\pi(u) := \mu(u_1 + u_2 + u_3 + \sigma g(u)).$$

We considered different forms for the misspecification $g(\cdot)$:

- quadratic effect: (a) $g(u) = 2u_1^2$, (b) $g(u) = 2u_5^2$,
- interactions:
  (c) $g(u) = u_1u_2$, (d) $g(u) = u_1u_3$, (f) $g(u) = u_1u_4$, (g) $g(u) = u_4u_7$.

Here $\sigma \geq 0$ measures the size of departure from the null hypothesis $H_0 : \sigma = 0$. Note that our GRP testing methodology requires an auxiliary sample of size n_A. We therefore randomly split the sample taking $n_A = n = N/2$, with n being the number of observations in the main sample. The observation vectors x_i follow a $\mathcal{N}_{10}(0, \Sigma_0)$ distribution where

$$(\Sigma_0)_{ij} := \rho^{|i-j|} \tag{14}$$

is the Toeplitz matrix with correlation parameter $\rho = 0.6$. The results for the six settings above are shown in Figure 3. All methods maintain good control over type I error, but in most scenarios our GRP-test has significantly greater power compared with the other methods.

4.2 High-dimensional settings

Here we consider logistic regression models with two different types of misspecification from the presence of a pure quadratic effect and an interaction term. Specifically, we take the log-odds to be

$$f_0(u) = u^T\beta_0 + \sigma g(u_1, \ldots, u_p)$$

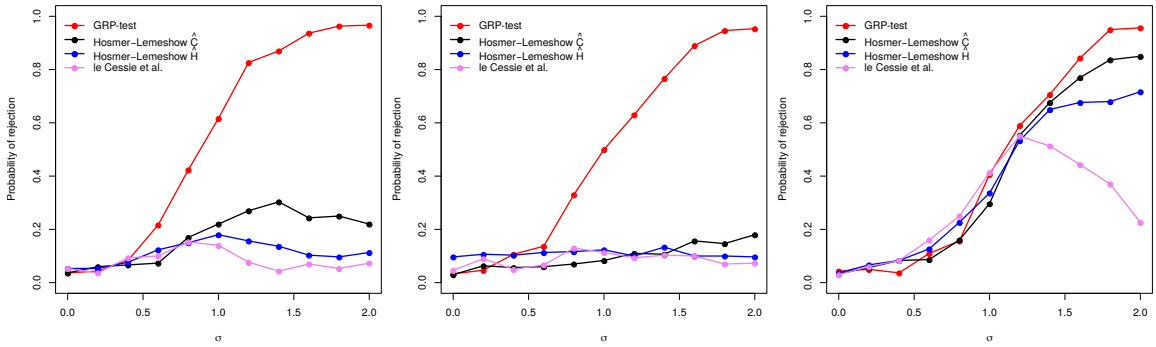
with

$$\beta_0 := (1, 1, 1, 1, 1, 0, \ldots, 0) \in \mathbb{R}^p$$

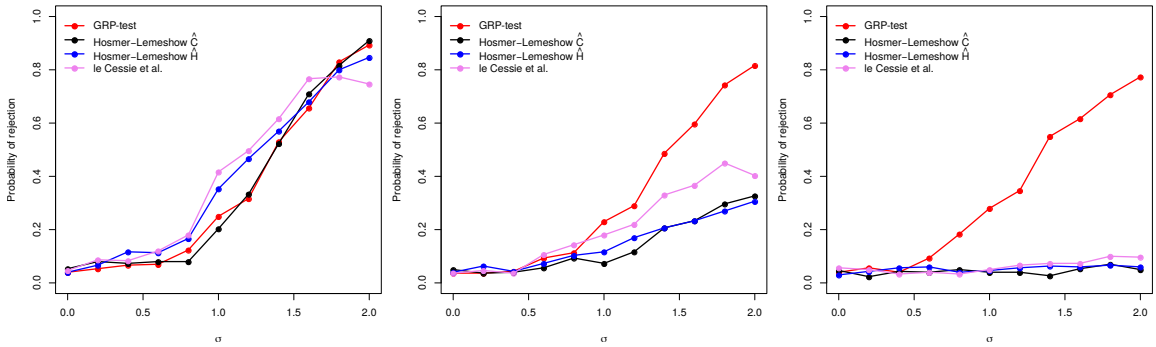
and either (a) $g(u) = u_1u_2$ or (b) $g(u) = u_1^2/2$. The parameter σ controls the degree of the misspecification and we look at $\sigma \in [0, 3]$.

The observation vectors x_i are independent with a Gaussian distribution $\mathcal{N}_p(0, \Sigma_0)$, where Σ_0 is the Toeplitz matrix (14) for a range of correlations $\rho \in \{0.4, 0.6, 0.8\}$. We consider two different settings on the dimension of the problem: (i) $p = 500, N = 800$, and (ii) $p = 3000, N = 2000$. The GRP-test requires sample splitting and we use split size 50%; thus the size of auxiliary sample is $n_A = 400$ and $n_A = 1000$, respectively. Again, we use a random forest as the residual prediction method.

In the high-dimensional setting, there is no obvious method that we can use for comparison with our proposed GRP-test. Therefore as a theoretical benchmark, we consider an oracle GRP-test applied to a reduced design matrix containing only variables in the active set $S = \{1, \ldots, 5\}$, thereby reducing problem to a low-dimensional one. The results are reported in Table 1 and Figure 2, from which we see that the GRP-test does indeed control the type I error, and suffers only a relatively small loss in power compared with the oracle GRP-test.



(a) Quadratic effect u_1^2 (b) Quadratic effect u_5^2 (c) Interaction u_1u_2



(d) Interaction u_1u_3 (e) Interaction u_1u_4 (f) Interaction u_4u_7

Figure 1: Comparison of GRP-test with Hosmer–Lemeshow $\hat{C}$, Hosmer–Lemeshow $\hat{H}$ and le Cessie–van Houwelingen–Copas–Hosmer unweighted sum of squares test.

Detecting the quadratic effect σu_1^2

GRP-test	$\sigma = 0$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$
$\rho = 0.4$	0.02	0.16	0.86	0.96
$\rho = 0.6$	0.04	0.18	0.82	0.94
$\rho = 0.8$	0.06	0.12	0.52	0.96

Benchmark	$\sigma = 0$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$
$\rho = 0.4$	0.05	0.52	0.99	1.00
$\rho = 0.6$	0.02	0.35	0.92	1.00
$\rho = 0.8$	0.05	0.18	0.76	0.99

Detecting the interaction effect σu_1u_2

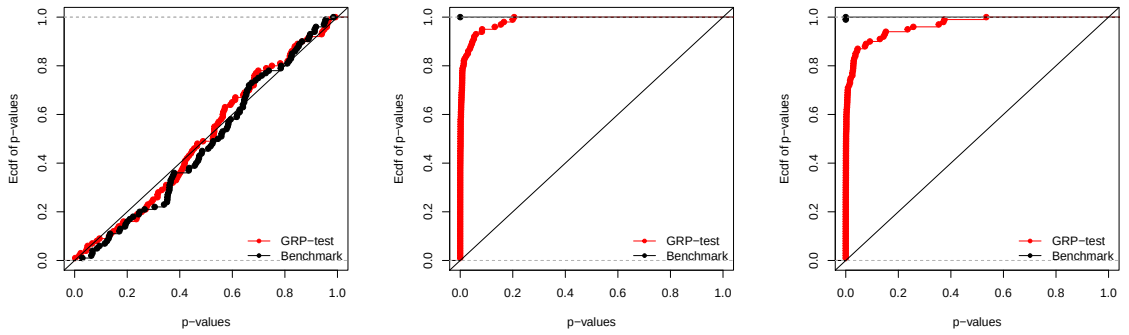
GRP-test	$\sigma = 0$	$\sigma = 1$	$\sigma = 2$	$\sigma = 3$
$\rho = 0.4$	0.02	0.16	0.86	0.94
$\rho = 0.6$	0.04	0.14	0.96	1.00
$\rho = 0.8$	0.06	0.34	1.00	1.00

Benchmark	$\sigma = 0$	$\sigma = 1$	$\sigma = 2$	$\sigma = 3$
$\rho = 0.4$	0.05	0.68	1.00	1.00
$\rho = 0.6$	0.04	0.70	1.00	1.00
$\rho = 0.8$	0.04	0.38	1.00	1.00

Table 1: Estimated probabilities of rejection of $H_0 : \sigma = 0$ at significance level 0.05 for different values of ρ and σ. Dimensions of the data are $p = 500$ and $N = 800$. Averages for the GRP-test were calculated over 50 iterations.

4.3 Semi-real data example

We use a gene-expression dataset on lung cancer available from the NCBI database (Spira et al. (2007), `https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2771`) to illustrate the size



(a) No misspecification (b) Quadratic effect σu_1^2 ($\sigma = 1.5$) (c) Interaction effect $\sigma u_1 u_2$ ($\sigma = 2$)

Figure 2: Empirical distribution functions of p-values for testing the goodness-of-fit in three scenarios. In this example, $p = 3000, N = 2000$.

and power performance of the goodness-of-fit test. We aim to detect if the model is a logistic regression, or if there are extra nonlinear effects. The full dataset contains airway epithelial gene expressions for 22215 genes from each of 192 smokers with (suspected) lung cancer, but this was reduced by taking the 500 genes with the largest variances. Having scaled the resulting variables, we fit a ℓ_1-penalized logistic regression using `cv.glmnet()` from the package `glmnet` (Friedman et al., 2010) and obtained a parameter estimate $\hat{\beta}$ with its corresponding support set $\hat{S}$. We then fit a Gaussian copula model to the rows of the design matrix and generated a new, augmented design matrix X by simulating a further $N = 800$ observation vectors from this fitted model. Finally, we generated 800 new responses: $Y_i | x_i = u \sim \text{Bern}\big(\hat{\pi}(u)\big)$ where

$$\hat{\pi}(u) := u^T \hat{\beta} + \sqrt{3} \frac{g(u) - \bar{g}}{\hat{\sigma}_g} \mathbb{1}_{\{\hat{\sigma}_g \neq 0\}},$$

$\bar{g} := N^{-1} \sum_{i=1}^N g(x_i)$ and $\hat{\sigma}_g^2 := (N-1)^{-1} \sum_{i=1}^N \{g(x_i) - \bar{g}\}^2$, for the following three scenarios:

$$\begin{aligned} g(u) &= 0, \\ g(u) &= u_{j_1}^2 + u_{j_2}^2, \\ g(u) &= u_{j_1} u_{j_2} + u_{j_3} u_{j_4}, \end{aligned}$$

where $j_\ell \in \hat{S}$, $\ell = 1, \ldots, 4$ are uniformly sampled entries from $\hat{S}$. We report rejection probabilities for all three scenarios from 100 repetitions in Table 2. In each case, the GRP-test is able to detect the misspecification relatively reliably, while keeping the type I error under control.

4.4 Group testing

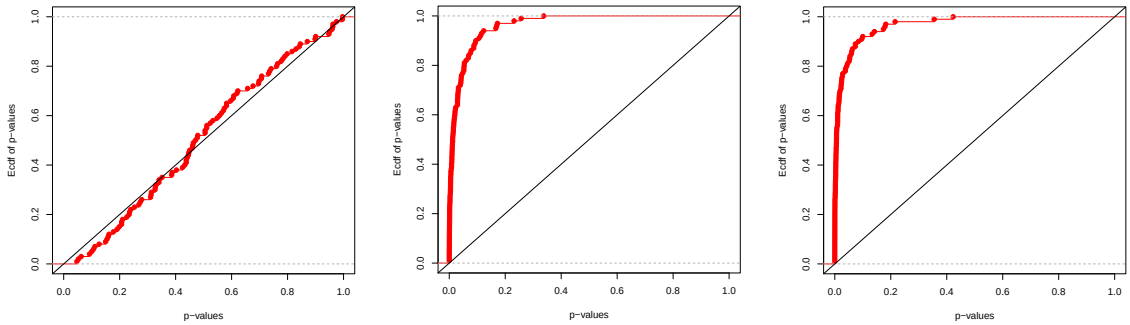
Finally, we consider the problem of testing for the significance of groups of predictors using the methodology set out in Section 3.1.2. We compare the GRP-test (Algorithm 2) with the globaltest

Testing goodness-of-fit of logistic regression on semi-real data on lung cancer

	Prob. of rejection of H_0
$g(u) = 0$	0.07
$g(u) = u_{j_1}^2 + u_{j_2}^2$	0.91
$g(u) = u_{j_1} u_{j_2} + u_{j_3} u_{j_4}$	0.74

Table 2: Estimated probabilities of rejection of $H_0 : g(u) = 0$ at significance level 0.05, averaged over 100 generated datasets.

Testing goodness-of-fit of logistic regression on semi-real data on lung cancer



(a) No misspecification: $g(u) = 0$. (b) Quadratic effects: $g(u) = u_{j_1}^2 + u_{j_2}^2$. (c) Interactions: $g(u) = u_{j_1} u_{j_2} + u_{j_3} u_{j_4}$.

Figure 3: Empirical distribution functions of p-values for testing the goodness-of-fit in the three scenarios.

(Goeman et al., 2004) and the de-biased Lasso (van de Geer et al., 2014; Dezeure et al., 2015) for logistic regression. We consider logistic regression models with coefficient vector of the form

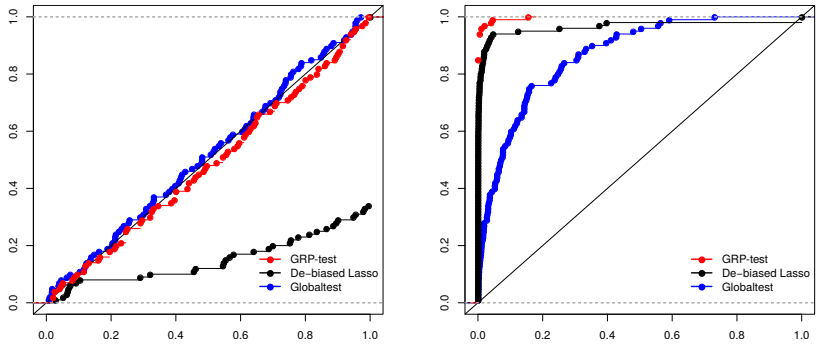
$$\beta_0 = (1, 1, 1, 1, \theta, 0, \ldots, 0) \in \mathbb{R}^p$$

for a range of values of $\theta \in [0, 1.5]$, and look at testing the null hypothesis $\beta_{0,G} = 0$ where $G = \{5, 6, \ldots, p\}$. Thus larger values of θ correspond to more extreme violations of the null. Similarly to earlier examples, we use design matrices constructed via realisations of a random Gaussian design with Toeplitz covariance (14) and $\rho = 0.6$. The results are reported in Figures 4 and 5. Both the GRP-test and the globaltest control the type I error at very close to the nominal level, while from Figure 4 the de-biased Lasso test is conservative; on the other hand, the GRP-test does very well in these examples in terms of power.

5 Discussion

In this work, we have introduced a new method for detecting conditional mean misspecification in generalized linear models based on predicting remaining signal in the residuals. For this task

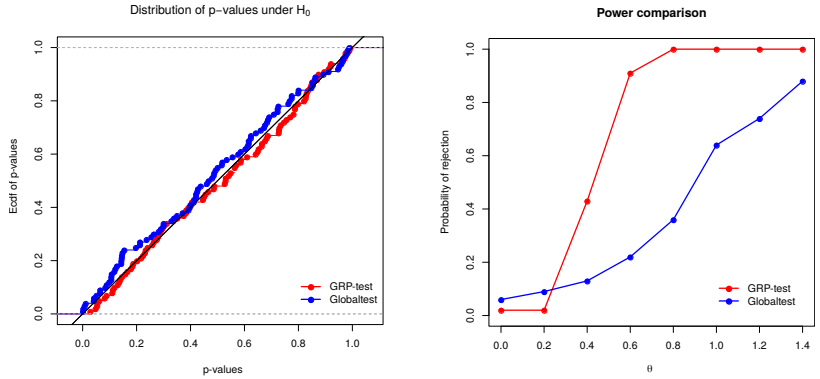
Group testing in logistic regression: comparison of GRP-test, de-biased Lasso and globaltest



(a) $\theta = 0$ (H_0 true) (b) $\theta = 1$ (H_0 false)

Figure 4: Comparison of GRP-test from Algorithm 2 with the de-biased Lasso (Dezeure et al., 2015) and globaltest (Goeman et al., 2004) when testing $H_0 : \beta_{0,G} = 0$, with $G = \{5, \ldots, p\}$, where $\beta_0 = (1, 1, 1, 1, \theta, 0, \ldots, 0)$. Plots show the empirical distribution functions of p-values under the null hypothesis (left) and under the alternative $\theta = 1$ (right). The dimensions of the data are $n = 500, p = 100$.

Group testing in logistic regression: comparison of GRP-test and globaltest



(a) $\theta = 0$ (H_0 true) (b) Powers for a range of θ (H_0 false)

Figure 5: Comparison of GRP-test from Algorithm 2 and globaltest (Goeman et al., 2004) when testing $H_0 : \beta_{0,G} = 0$, with $G = \{5, \ldots, 100\}$, where $\beta_0 = (1, 1, 1, 1, \theta, 0, \ldots, 0)$. Plots show the empirical distribution functions of p-values under the null hypothesis (left) and powers of the test at significance level $\alpha = 0.05$ for a range of values of θ which are on the x-axis (right). The dimensions of the data are $p = 800, n = 600$. The de-biased Lasso test is omitted since it is very computationally expensive.

of prediction, we have a number of powerful machine learning methods at our disposal. Whilst these estimation performance of these methods is largely theoretically intractable, by employing sample-splitting and a careful debiasing strategy involving the square-root Lasso, our generalized

residual prediction framework provides formal statistical tests with type I error control when used in conjunction with (essentially) arbitrary machine learning methods.

One requirement for these theoretical guarantees is that the sparsity of the true regression coefficient β_0 satisfies $s = o(\sqrt{n}/\log p)$, a condition that was also needed in related work on the de-biased Lasso (Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014; Cai and Guo, 2017). It would be very interesting if this could be relaxed to $s = o(n/\log p)$ for instance, which would encompass settings where the GLM Lasso estimate satisfies $\|\hat{\beta} - \beta_0\|_2 \to 0$ though $\|\hat{\beta} - \beta_0\|_1$ may be diverging.

Another interesting question is whether sample splitting can be completely avoided if we were able to obtain guarantees for an estimator $\hat{w}$ of a population direction w. Such alternatives to sample splitting could be particularly helpful for settings where there is dependence across the observations, such as in the case of generalized linear mixed effect models.

6 Proofs

6.1 Proofs for Section 3.1.1

Proof of Theorem 1. This follows from the more general Theorem 3, noting that under the null hypothesis, $\mathbf{\Delta} = 0$. The proof of Theorem 3 can be found in Section 6.3. □

6.2 Proofs for Section 3.1.2

Proof of Proposition 1. For $j \in G$, let $u_j := X_j - X_{-G}\gamma_{0,j}$ and define the sets

$$\mathcal{T}_{1,j} := \left\{ \|u_j^T D_{\beta_0} X_{-G}\|_\infty / n \leq 6C_0K^2\sqrt{\log(2p)/n} \right\},$$

$$\mathcal{T}_2 := \left\{ \hat{\beta}_{-G} \in \Theta_{-G}(\lambda, \beta_0) \right\},$$

$$\mathcal{T}_{3,j} := \left\{ \|X_{j\cup G^c}^T (Y - \mu(X\beta_0))\|_\infty / n \leq AK\sqrt{\log p/n} \right\},$$

where A will be specified below. We first derive a high-probability bound for the set

$$\mathcal{T} := \cap_{j\in G}(\mathcal{T}_{1,j} \cap \mathcal{T}_2 \cap \mathcal{T}_{3,j}).$$

By Lemma 3 in Section A, there exists a constant $A > 0$ (in the definition of $\mathcal{T}_{3,j}$) such that

$$\mathbb{P}(\mathcal{T}^c) \leq 3/p. \tag{15}$$

In the rest of the proof, we work on the event $\mathcal{T}$. Define $D^2_{\beta_{0,-G}} := \mu'(X_{-G}\beta_{0,-G})$ (note that under H_0, it holds that $D_{\beta_{0,-G}} = D_{\beta_0}$). Now consider the decomposition

$$\begin{aligned} T_j = \hat{w}_j^T D^{-1}_{\hat{\beta}_{-G}} \big(Y - \mu(X_{-G}\hat{\beta}_{-G})\big) &= w_j^T D^{-1}_{\beta_{0,-G}} \big(Y - \mu(X_{-G}\beta_{0,-G})\big) + \mathrm{rem}_{1,j} + \mathrm{rem}_{2,j} \\ &= w_j^T \varepsilon + \mathrm{rem}_{1,j} + \mathrm{rem}_{2,j}, \end{aligned} \tag{16}$$

where

$$\text{rem}_{1,j} := (D^{-1}_{\hat{\beta}_{-G}}\hat{w}_j - D^{-1}_{\beta_{0,-G}}w_j)^T\big(Y - \mu(X_{-G}\beta_{0,-G})\big),$$
$$\text{rem}_{2,j} := \hat{w}_j^T D^{-1}_{\hat{\beta}_{-G}}\big(\mu(X_{-G}\beta_{0,-G}) - \mu(X_{-G}\hat{\beta}_{-G})\big),$$

and where we write $D_{\hat{\beta}_{-G}} := \hat{D}$. We first derive the rates of convergence for the estimator $\hat{\gamma}_j$ from Algorithm 2 and then proceed to bound the remainders. By Lemma 4 in Section A, there exist positive constants C_3 and C_4 such that on $\mathcal{T}$, we have

$$\max_{j\in G}\|\hat{\gamma}_j - \gamma_{0,j}\|_1 \le C_3K^2 s\sqrt{\log p/n}, \tag{17}$$
$$\max_{j\in G}|\hat{\tau}_j^2 - \tau_j^2| \le C_4K^2\sqrt{s\log p/n}, \tag{18}$$

where $\hat{\tau}_j^2 := \big\|D_{\hat{\beta}_{-G}}(X_j - X_{-G}\hat{\gamma}_j)\big\|_2^2/n$.

Remainder $\text{rem}_{1,j}$: Let k_j denote the index of column X_j in the matrix $X_{j\cup G^c}$. For $j \in G$, we define

$$\hat{\Gamma}_j := (-\hat{\gamma}_{j,1}, \ldots, -\hat{\gamma}_{j,k_j-1}, 1, -\hat{\gamma}_{j,k_j+1}, \ldots, \hat{\gamma}_{j,|G^c|})^T$$

and its population-level counterpart $\Gamma_{0,j}$ based on $\gamma_{0,j}$. First note that

$$D^{-1}_{\hat{\beta}_{-G}}\hat{w}_j = (X_j - X_{-G}\hat{\gamma}_j)/(\sqrt{n}\hat{\tau}_j) = X_{j\cup G^c}\hat{\Gamma}_j/(\sqrt{n}\hat{\tau}_j),$$

and similarly,

$$D^{-1}_{\beta_{0,-G}}w_j = (X_j - X_{-G}\gamma_{0,j})/(\sqrt{n}\tau_j) = X_{j\cup G^c}\Gamma_{0,j}/(\sqrt{n}\tau_j). \tag{19}$$

Therefore, we obtain using Hölder's inequality that

$$\begin{aligned}|\text{rem}_{1,j}| &= \big|(\hat{\Gamma}_j/\hat{\tau}_j - \Gamma_{0,j}/\tau_j)^T X^T_{j\cup G^c}(Y - \mu(X\beta_0))\big|/\sqrt{n}\\ &\le \|\hat{\Gamma}_j/\hat{\tau}_j - \Gamma_{0,j}/\tau_j\|_1\big\|X^T_{j\cup G^c}(Y - \mu(X\beta_0))\big\|_\infty/\sqrt{n}.\end{aligned}$$

On the set $\cap_{j\in G}\mathcal{T}_{3,j} \subseteq \mathcal{T}$, we have

$$\max_{j\in G}|\text{rem}_{1,j}| \le \max_{j\in G}\|\hat{\Gamma}_j/\hat{\tau}_j - \Gamma_{0,j}/\tau_j\|_1 AK\sqrt{\log p}. \tag{20}$$

Next we bound $\|\hat{\Gamma}_j/\hat{\tau}_j - \Gamma_{0,j}/\tau_j\|_1$. Firstly, we can decompose and bound

$$\begin{aligned}\|\hat{\Gamma}_j/\hat{\tau}_j - \Gamma_{0,j}/\tau_j\|_1 &= \|(\hat{\Gamma}_j - \Gamma_{0,j})/\hat{\tau}_j + \Gamma_{0,j}(1/\hat{\tau}_j - 1/\tau_j)\|_1\\ &\le \|\hat{\gamma}_j - \gamma_{0,j}\|_1/\hat{\tau}_j + \|\Gamma_{0,j}\|_1|1/\hat{\tau}_j - 1/\tau_j|.\end{aligned} \tag{21}$$

Now note that

$$\begin{aligned}\tau_j^2 &= \mathbb{E}(X_j - X_{-G}\gamma_{0,j})^T D^2_{\beta_0}(X_j - X_{-G}\gamma_{0,j})/n\\ &= \mathbb{E}X_j^T D^2_{\beta_0}(X_j - X_{-G}\gamma_{0,j})/n\\ &= \mathbb{E}X_j^T D^2_{\beta_0}X_{j\cup G^c}\Gamma_{0,j}/n.\end{aligned} \tag{22}$$

Combining (22) with the fact that

$$\mathbb{E}X_{-G}^T D_{\beta_0}^2 X_{j\cup G^c}\Gamma_{0,j}/n = 0,$$

we obtain that

$$\mathbb{E}X_{j\cup G^c}^T D_{\beta_0}^2 X_{j\cup G^c}\Gamma_{0,j}/n = \tau_j^2 e_{k_j},$$

where e_ℓ denotes the ℓ-th standard basis vector in $\mathbb{R}^{p-|G|+1}$. Define $\Sigma_{\beta_0,j\cup G^c} := \mathbb{E}X_{j\cup G^c}^T D_{\beta_0}^2 X_{j\cup G^c}/n$ and note from Condition 2(iv) that $\Sigma_{\beta_0,j\cup G^c}$ is invertible. Consequently, $\Gamma_{0,j}/\tau_j^2 = \Sigma_{\beta_0,j\cup G^c}^{-1} e_{k_j}$, so $1/\tau_j^2 = (\Sigma_{\beta_0,j\cup G^c}^{-1})_{k_j k_j}$. Further note that

$$\max_{j\in G}\tau_j^2 = \max_{j\in G} 1/(\Sigma_{\beta_0,j\cup G^c}^{-1})_{k_j k_j} \le \max_{j\in G}(\Sigma_{\beta_0,j\cup G^c})_{k_j k_j} \le \|\Sigma_0\|_\infty \le C_e.$$

Therefore, by Condition 2(iv), we have

$$\max_{j\in G}\|\Gamma_{0,j}\|_2 = \max_{j\in G}\|\Sigma_{\beta_0,j\cup G^c}^{-1} e_j\|_2\tau_j^2 \le \max_{j\in G}\Lambda_{\min}^{-1}(\Sigma_{\beta_0,j\cup G^c})\tau_j^2 \le \Lambda_{\min}^{-1}(\Sigma_0)\max_{j\in G}\tau_j^2 \le C_e^2.$$

Consequently, and by sparsity of $\gamma_{0,j}$ assumed in Condition 2(v), it follows that

$$\max_{j\in G}\|\Gamma_{0,j}\|_1 \le \sqrt{s+1}\max_{j\in G}\|\Gamma_{0,j}\|_2 \le C_e^2\sqrt{s+1}.$$

Moreover,

$$\begin{aligned}\min_{j\in G}\tau_j^2 &= \min_{j\in G}\frac{1}{\big(\Sigma_{\beta_0,j\cup G^c}^{-1}\big)_{k_j k_j}} \ge \min_{j\in G}\frac{1}{\big\|\Sigma_{\beta_0,j\cup G^c}^{-1}\big\|_{\mathrm{op}}} = \min_{j\in G}\big\|\Sigma_{\beta_0,j\cup G^c}\big\|_{\mathrm{op}} \\ &\ge \min_{j\in G}\Lambda_{\min}\big(\Sigma_{\beta_0,j\cup G^c}\big) \ge \Lambda_{\min}\big(\Sigma_0\big) \ge \frac{1}{C_e}. \end{aligned} \tag{23}$$

By Condition 2(v), we can find $n_0 \in \mathbb{N}$ such that $a_n \le 1/(2C_4C_e)$ for $n \ge n_0$. Then from (18) and (23), we obtain on $\mathcal{T}$ that for $n \ge n_0$,

$$\min_{j\in G}\frac{\hat\tau_j^2}{\tau_j^2} \ge 1 - \max_{j\in G}\frac{|\hat\tau_j^2 - \tau_j^2|}{\tau_j^2} \ge 1 - C_4C_eK^2\sqrt{s\log p/n} \ge \frac{1}{2}.$$

Using (21), we conclude that on $\mathcal{T}$,

$$\max_{j\in G}\|\hat\Gamma_j/\hat\tau_j - \Gamma_{0,j}/\tau_j\|_1 \le C_eC_3K^2 s\sqrt{\log p/n} + C_e^2C_4K^2\sqrt{s\log p/n}. \tag{24}$$

Consequently, combining (20) and (24), there exists a constant such that it holds on $\mathcal{T}$

$$\max_{j\in G}|\mathrm{rem}_{1,j}| = \mathcal{O}\big(K^3 s\log p/\sqrt{n}\big) = \mathcal{O}(a_n). \tag{25}$$

Remainder $\mathbf{rem}_{2,j}$: By the mean value theorem, for each $i = 1,\dots,n$, there exists $\alpha_i \in [0,1]$ such that

$$\mu(x_{i,-G}^T\beta_{0,-G}) - \mu(x_{i,-G}^T\hat\beta_{-G}) = \mu'(x_{i,-G}^T\tilde\beta_{(i)})x_{i,-G}^T(\beta_{0,-G} - \hat\beta_{-G}),$$

where $\tilde{\beta}_{(i)} := \alpha_i \hat{\beta}_{-G} + (1-\alpha_i)\beta_{0,-G}$. Consequently, using Hölder's inequality and the KKT conditions from the optimization problem in Algorithm 2, we obtain

$$\begin{aligned}
|\mathrm{rem}_{2,j}| &= \sum_{i=1}^{n} \big| \hat{w}_{j,i} D^{-1}_{\hat{\beta}_{-G},ii} \mu'(X^T_{i,-G}\tilde{\beta}_{(i)}) X^T_{i,-G}(\beta_{0,-G} - \hat{\beta}_{-G}) \big| \\
&\le \big| \hat{w}_j^T D^{-1}_{\hat{\beta}_{-G}} D^2_{\hat{\beta}_{-G}} X_{-G}(\beta_{0,-G} - \hat{\beta}_{-G}) \big| + \mathrm{rem}_{3,j} \\
&\le \|\hat{w}_j^T D_{\hat{\beta}_{-G}} X_{-G}\|_\infty \|\beta_{0,-G} - \hat{\beta}_{-G}\|_1 + \mathrm{rem}_{3,j} \\
&= \mathcal{O}\big(K\sqrt{\log p}\, s\sqrt{\log p/n}\big) + \mathrm{rem}_{3,j}, \qquad (26)
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{rem}_{3,j} &:= \left| \sum_{i=1}^{n} \hat{w}_{j,i} D^{-1}_{\hat{\beta}_{-G},ii} \big(\mu'(x^T_{i,-G}\tilde{\beta}_{(i)}) - \mu'(x^T_{i,-G}\hat{\beta}_G)\big) x^T_{i,-G}(\beta_{0,-G} - \hat{\beta}_{-G}) \right| \\
&\le L \sum_{i=1}^{n} |\hat{w}_{j,i}| \big| D^{-1}_{\hat{\beta}_{-G},ii} \big| \big\{ x^T_{i,-G}(\beta_{0,-G} - \hat{\beta}_{-G}) \big\}^2.
\end{aligned}$$

By Condition 2(ii), we have $\|X_{-G}\gamma_{0,j}\|_\infty \le K$, so we obtain

$$\begin{aligned}
\max_{j\in G} \|w_j\|_\infty &= \max_{j\in G} \|D_{\beta_0}(X_j - X_{-j}\gamma_{0,j})\|_\infty / (\sqrt{n}\tau_j) \\
&\le \max_{j\in G} \frac{1}{\sqrt{n}\tau_j} \big\{ \|D_{\beta_0} X_j\|_\infty + \|D_{\beta_0} X_{-G}\gamma_{0,j}\|_\infty \big\} \\
&\le \frac{2C_e C_0^{1/2} K}{\sqrt{n}}. \qquad (27)
\end{aligned}$$

Since μ' is Lipschitz and $\|x_i\|_\infty \le K$, we obtain that on $\mathcal{T}$,

$$\big| D^2_{\hat{\beta}_{-G},ii} - D^2_{\beta_0,ii} \big| \le L |x^T_{i,-G}(\hat{\beta}_{-G} - \beta_{0,-G})| \le L\|x_i\|_\infty \|\hat{\beta}_{-G} - \beta_{0,-G}\|_1 \le LKs\lambda. \qquad (28)$$

Thus

$$\max_{i=1,\dots,n} \big| D_{\hat{\beta}_{-G},ii} \big| \asymp 1 \qquad (29)$$

by Condition 2(iii) and Condition 2(v). Therefore, on $\mathcal{T}$, it follows that

$$\begin{aligned}
\big\| D_{\hat{\beta}_{-G}}(X_j - X_{-G}\hat{\gamma}_j) \big\|_\infty &\le \big\| D_{\hat{\beta}_{-G}}(X_j - X_{-G}\gamma_{0,j}) \big\|_\infty + \big\| D_{\hat{\beta}_{-G}} X_{-G}(\hat{\gamma}_j - \gamma_{0,j}) \big\|_\infty \\
&\le \mathcal{O}(K) + \|D_{\hat{\beta}_{-G}} X_{-G}\|_\infty \|\hat{\gamma}_j - \gamma_{0,j}\|_1 \\
&= \mathcal{O}(K) + \mathcal{O}(K^3 s\sqrt{\log p/n}) = \mathcal{O}(K).
\end{aligned}$$

Consequently, on $\mathcal{T}$,

$$\max_{j\in G} \|\hat{w}_j\|_\infty = \max_{j\in G} \|D_{\hat{\beta}_{-G}}(X_j - X_{-G}\hat{\gamma}_j)\|_\infty / (\sqrt{n}\hat{\tau}_j) = \mathcal{O}\left(\frac{K}{\sqrt{n}}\right). \qquad (30)$$

Then using the result above, on $\mathcal{T}$, we obtain

$$\begin{aligned}\mathrm{rem}_{3,j} &\leq L\sum_{i=1}^{n}|\hat{w}_{j,i}|\big|D^{-1}_{\hat{\beta}_{-G},ii}\big|\big\{x^T_{i,-G}(\beta_{0,-G}-\hat{\beta}_{-G})\big\}^2\\ &= \mathcal{O}\big(K\sqrt{n}s\lambda^2\big)=\mathcal{O}(a_n).\end{aligned} \tag{31}$$

The result follows from (16), together with the bounds (15), (25), (26) and (31). □

Proof of Theorem 2. We want to show that the quantiles of our test statistic for group testing,

$$T := \max_{j\in G}\left|\sum_{i=1}^{n}\hat{w}_{j,i}\hat{R}_{G,i}\right|$$

can be approximated by quantiles of its bootstrapped version

$$W := T^b = \max_{j\in G}\left|\sum_{i=1}^{n}\hat{w}_{j,i}\hat{R}_{G,i}e^b_i\right|,$$

where $(e^b_i)_{i=1}^n$ is a sequence of independent and identically distributed $\mathcal{N}(0,1)$ random variables. We can apply Theorem 4 from Section A.3 together with Proposition 1. Adopting the notation of Theorem 4 we let

$$T_0 := \max_{j\in G}\left|\sum_{i=1}^{n}w_{j,i}\varepsilon_i\right|,$$

and

$$W_0 := \max_{j\in G}\left|\sum_{i=1}^{n}w_{j,i}\varepsilon_i e^b_i\right|.$$

Note that the maxima above can be rewritten without the absolute values using the fact that for any $a\in\mathbb{R}$ it holds that $|a|=\max\{a,-a\}$. Thus for $i=1,\dots,n$ and $j\in G$ we let $x_{ij}:=\sqrt{n}w_{j,i}\varepsilon_i$ and $\hat{x}_{ij}:=\sqrt{n}\hat{w}_{j,i}\hat{R}_{G,i}$. Moreover, for $i=1,\dots,n$ and $j\in G$ we also define $x_{i(j+|G|)}:=-x_{ij}$ and $\hat{x}_{i(j+|G|)}:=-\hat{x}_{ij}$. We will apply Theorem 4 with x_{ij} and $\hat{x}_{ij}$ where $i=1,\dots,n$ and $j\in H:=G\cup\{j+|G|:j\in G\}$.

We now check that conditions (53), (54), (55), (56) and (57) needed for Theorem 4 are satisfied.

Checking condition (53):

First, by the tower property, we have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}x^2_{ij}=\sum_{i=1}^{n}\mathbb{E}\big\{w^2_{j,i}\mathbb{E}(\varepsilon^2_i|X)\big\}.$$

By Condition 2(iii), it follows that $c_0\leq D^2_{\beta_0,ii}\leq C_0$. Consequently, and using Condition 2(i), there exist constants c,C such that $c\leq\mathbb{E}(\varepsilon^2_i|X)=\mathbb{E}(\eta^2_i/D^2_{\beta_0,ii}|X)\leq C$. It follows that

$$c\sum_{i=1}^{n}\mathbb{E}w^2_{j,i}\leq\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}x^2_{ij}=\frac{1}{n}\sum_{i=1}^{n}n\mathbb{E}\big\{w^2_{j,i}\mathbb{E}(\varepsilon^2_i|X)\big\}\leq C\sum_{i=1}^{n}\mathbb{E}w^2_{j,i}. \tag{32}$$

Now recalling that

$$w_j = D_{\beta_0}(X_j - X_{-G}\gamma_{0,j})/(\sqrt{n}\tau_j),$$

we see that

$$\sum_{i=1}^{n} \mathbb{E}w_{j,i}^2 = \mathbb{E}\|w_j\|_2^2 = 1. \tag{33}$$

Therefore, combining (32) and (33), we obtain $c \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}x_{ij}^2 \leq C$ for $j \in H$, as required.

Checking condition (54):

Recall from (27) that we have the deterministic bound

$$\max_{i,j}|w_{j,i}| = \mathcal{O}(K/\sqrt{n}).$$

Using this bound, we will now check that for suitable $B_n \asymp K$,

$$\max_{k=1,2}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|x_{ij}|^{2+k}/B_n^k + \mathbb{E}\max_{j\in H}|x_{ij}/B_n|^4 \leq 4. \tag{34}$$

First observe that

$$\begin{aligned}
\max_{k=1,2}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|x_{ij}|^{2+k}/B_n^k &\lesssim \max_{k=1,2} n^{k/2}\sum_{i=1}^{n}\mathbb{E}w_{j,i}^{2+k}/B_n^k \\
&\lesssim \max_{k=1,2} n^{k/2}\left(\frac{K}{\sqrt{n}}\right)^k\sum_{i=1}^{n}\mathbb{E}w_{j,i}^2/B_n^k \\
&\lesssim \max_{k=1,2}(K/B_n)^k,
\end{aligned}$$

and

$$\mathbb{E}\max_{j\in H}|x_{ij}/B_n|^4 \lesssim (K/B_n)^4.$$

Taking sufficiently large $B_n \asymp K$, we can therefore guarantee that (34) holds, as required.

Checking condition (55):

By Proposition 1, there exists a constant $C' > 0$ such that

$$\mathbb{P}(|T - T_0| > \zeta_1) \leq \mathbb{P}\left(\max_{j\in H}|T_j - w_j^T\varepsilon| > \zeta_1'\right) \leq \zeta_2',$$

for $\zeta_1' := C'K^3 s\log p/\sqrt{n}$ and $\zeta_2' := 3/p$. Next note that

$$|W - W_0| \leq \max_{j\in H}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_{ij} - \hat{x}_{ij})e_i^b\right|.$$

Now conditional on $(x_{ij})_{i=1}^n$ and $(\hat{x}_{ij})_{i=1}^n$ we have that $Z_j := \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_{ij}-\hat{x}_{ij})e_i^b \sim \mathcal{N}(0,\sigma_j^2)$ where $\sigma_j^2 := \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \hat{x}_{ij})^2$. Therefore,

$$\mathbb{E}_e|W - W_0| \leq \mathbb{E}_e\max_{j\in H}|Z_j| \leq \sqrt{2\log(|H|+1)}\max_{j\in H}\sigma_j.$$

Then it follows by Borell's inequality that for any $t > 0$,

$$\mathbb{P}_e\Big(|W - W_0| > t + \mathbb{E}_e \max_{j\in H}|Z_j|\Big) \le \mathbb{P}_e\Big(\max_{j\in H}|Z_j| > t + \mathbb{E}_e \max_{j\in H}|Z_j|\Big) \le e^{-t^2/(2\max_j \sigma_j^2)}.$$

Taking $t := \sqrt{2\log(2p+1)}\max_{j\in H}\sigma_j$ and noting that $|H| = 2|G| \le 2p$, we obtain

$$\mathbb{P}_e\left(|W - W_0| > 2\sqrt{2\log(2p+1)}\max_{j\in H}\sigma_j\right) \le e^{-\log(2p+1)} \le 1/p. \tag{35}$$

Denote $\Delta_2 := \max_{j\in H}\sigma_j^2$. Then by Lemma 5 in Appendix A there exists a constant $C'' > 0$ such that

$$\mathbb{P}\big(\Delta_2 \ge (C'')^2 K^6 s^2 \lambda^2\big) \le 4/p. \tag{36}$$

Therefore, combining (35) and (36), we obtain

$$\mathbb{P}\left(\mathbb{P}_e(|W - W_0| > \sqrt{2\log(2p)}C''K^3 s\lambda) > 1/p\right) \le 4/p.$$

Thus we can take

$$\zeta_1 := \max\{\sqrt{2\log(2p)}C''K^3 s\lambda,\ \zeta_1'\} \asymp K^3 s\log p/\sqrt{n},$$

and $\zeta_2 := 4/p$ (in applying Theorem 4).

Checking conditions (56) and (57):

Finally, by assumption (10), there exist constants $C_2', c_2 > 0$ such that

$$\zeta_1 \log(2|G|) + \zeta_2 \le C_2' n^{-c_2},$$

and

$$B_n^4 \log(2|G|n)^7/n \le C_2' n^{-c_2},$$

where $B_n \asymp K$. □

6.3 Proofs for Section 3.2

Proof of Theorem 3. In this proof, it is convenient to write $\mathbb{P}_A(\cdot)$ and $\mathbb{E}_A(\cdot)$ as shorthand for $\mathbb{P}(\cdot|Z_A)$ and $\mathbb{E}(\cdot|Z_A)$ respectively. Consider the decomposition

$$T = \hat{w}_A^T \hat{R} = \hat{w}_A^T \hat{D}_A^{-1}\big(Y - \mu(X\hat{\beta})\big) = \phi + \mathbf{\Delta} + \mathrm{rem}_1,$$

where

$$\begin{aligned}
\phi &:= \hat{w}_A^T \hat{D}_A^{-1}\big\{Y - \mu\big(f_0(X)\big)\big\},\\
\mathbf{\Delta} &:= \hat{w}_A^T \hat{D}_A^{-1}\big\{\mu\big(f_0(X)\big) - \mu(X\beta_0)\big\},\\
\mathrm{rem}_1 &:= \hat{w}_A^T \hat{D}_A^{-1}\big\{\mu(X\beta_0) - \mu(X\hat{\beta})\big\}.
\end{aligned}$$

There are three terms:

I. The term ϕ is the pivot. By the Berry–Esseen theorem, we will show below that (after scaling) it is well approximated by a normal random variable.

II. The term $\boldsymbol{\Delta}$ captures the deviation from the null hypothesis. If the null hypothesis is true, then $\boldsymbol{\Delta} = 0$.

III. The term rem_1 is a stochastic remainder term, for which we will develop a probabilistic bound below.

Let $\sigma := \|D_Y \hat{D}_A^{-1} \hat{w}_A\|_2$. Then

$$\sup_{z\in\mathbb{R}} |\mathbb{P}_A(\phi + \text{rem}_1 \le z) - \Phi(z)| \le \sup_{z\in\mathbb{R}} |\mathbb{P}_A(\phi + \text{rem}_1 \le z) - \Phi(z/\sigma)| + \sup_{z\in\mathbb{R}} |\Phi(z/\sigma) - \Phi(z)|. \quad (37)$$

Now, for any $\epsilon > 0$ and $z \in \mathbb{R}$,

$$\begin{aligned}\mathbb{P}_A(\phi + \text{rem}_1 \le z) &\le \mathbb{P}_A(\phi \le z + \epsilon) + \mathbb{P}_A(|\text{rem}_1| > \epsilon) \\ &\le \Phi(z/\sigma) + \sup_{x\in\mathbb{R}} |\mathbb{P}_A(\phi \le x) - \Phi(x/\sigma)| + \frac{\epsilon}{\sqrt{2\pi}\sigma} + \mathbb{P}_A(|\text{rem}_1| > \epsilon). \quad (38)\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{P}_A(\phi + \text{rem}_1 \le z) &\ge \mathbb{P}_A(\phi \le z - \epsilon) - \mathbb{P}_A(|\text{rem}_1| > \epsilon) \\ &\ge \Phi(z/\sigma) - \sup_{x\in\mathbb{R}} |\mathbb{P}_A(\phi \le x) - \Phi(x/\sigma)| - \frac{\epsilon}{\sqrt{2\pi}\sigma} - \mathbb{P}_A(|\text{rem}_1| > \epsilon). \quad (39)\end{aligned}$$

Therefore, combining (37), (38) and (39), we find that

$$\sup_{z\in\mathbb{R}} |\mathbb{P}_A\left(T - \boldsymbol{\Delta} \le z\right) - \Phi(z)| \le \sup_{x\in\mathbb{R}} |\mathbb{P}_A(\phi \le x) - \Phi(x/\sigma)| + \frac{\epsilon}{\sqrt{2\pi}\sigma} + \mathbb{P}_A(|\text{rem}_1| > \epsilon) + |\text{rem}_0|, \quad (40)$$

where $\text{rem}_0 := \frac{1}{\sqrt{2\pi}}\left(\frac{1}{\sigma} - 1\right)$.

Bound for the pivot.

We apply the Berry–Esseen theorem for non-identically distributed summands to $Z_0 := \phi/\sigma$. Note that

$$Z_0 = \frac{\sum_{i=1}^n \hat{w}_{A,i} \hat{D}_{A,ii}^{-1} \{Y_i - \mu(f_0(x_i))\}}{\sqrt{\sum_{i=1}^n \hat{w}_{A,i}^2 \hat{D}_{A,ii}^{-2} D_{Y,ii}^2}}.$$

For $i = 1, \ldots, n$, denote $U_i := \hat{w}_{A,i} \hat{D}_{A,ii}^{-1} \{Y_i - \mu(f_0(x_i))\}$ and $\sigma_i^2 := \text{Var}(U_i | Z_A) = \hat{w}_{A,i}^2 \hat{D}_{A,ii}^{-2} D_{Y,ii}^2$. Since $\mathbb{E}_A(U_i) = 0$, the Berry–Esseen theorem (Esseen, 1942) yields that

$$\sup_{x\in\mathbb{R}} |\mathbb{P}_A(\phi \le x) - \Phi(x/\sigma)| = \sup_{x\in\mathbb{R}} |\mathbb{P}_A(Z_0 \le x) - \Phi(x)| \le C_1 \frac{\sum_{i=1}^n \mathbb{E}_A(|U_i|^3)}{\{\sum_{i=1}^n \sigma_i^2\}^{3/2}},$$

where $C_1 > 0$ is a universal constant. Hence using Condition 1,

$$\sup_{x\in\mathbb{R}} |\mathbb{P}_A(\phi \le x) - \Phi(x/\sigma)| \le \frac{C_1 C_\varepsilon}{\sigma^3} \sum_{i=1}^n |\hat{w}_{A,i} D_{Y,ii} \hat{D}_{A,ii}^{-1}|^3 \le \frac{C_1 C_\varepsilon}{\sigma} \|D_Y \hat{D}_A^{-1} \hat{w}_A\|_\infty. \quad (41)$$

Bound for rem_0. To bound $|\sigma - 1|$, we first bound $\|D_Y^2 \hat{D}_A^{-2} - I\|_\infty$. By Hölder's inequality and using the assumptions $\hat{\beta}_A \in \Theta(\lambda_A, \beta_0, X_A)$ and $\|x_i\|_\infty \leq K_X$, we have

$$|x_i^T(\hat{\beta}_A - \beta_0)| \leq \|x_i\|_\infty \|\hat{\beta}_A - \beta_0\|_1 \leq K_X s \lambda_A. \tag{42}$$

Using the assumptions $|x_i^T \beta_0| \leq K_0/2$ and $K_X s\lambda_A \leq K_0/2$ together with (42), we obtain

$$|x_i^T \hat{\beta}_A| \leq |x_i^T \beta_0| + |x_i^T(\hat{\beta}_A - \beta_0)| \leq K_0/2 + K_X s\lambda_A \leq K_0/2 + K_0/2 = K_0.$$

Hence $x_i^T \hat{\beta}_A \in [-K_0, K_0]$. Then by the assumption that $V \circ \mu$ is L-Lipschitz on the interval $[-K_0, K_0]$ and using (42), we have

$$|\hat{D}_{A,ii}^2 - D_{\beta_0,ii}^2| = \left|V(\mu(x_i^T \hat{\beta}_A)) - V(\mu(x_i^T \beta_0))\right| \leq L|x_i^T(\hat{\beta}_A - \beta_0)| \leq LK_X s\lambda_A.$$

Therefore, $\hat{D}_{A,ii}^2 \geq D_{\beta_0,ii}^2/2$ under the condition $6d_{\min}^{-2} LK_X s\lambda_A \leq 1/2$. This then implies that

$$|\eta_{i,2}| := |D_{\beta_0,ii}^2/\hat{D}_{A,ii}^2 - 1| \leq 2d_{\min}^{-2} LK_X s\lambda_A.$$

Next, by assumption, we have

$$|\eta_{i,1}| := |D_{Y,ii}^2 D_{\beta_0,ii}^{-2} - 1| \leq 2d_{\min}^{-2} LK_X s\lambda_A,$$

(note that under H_0, it holds that $D_Y = D_{\beta_0}$, so $|\eta_{i,1}| = |D_{Y,ii}^2 D_{\beta_0,ii}^{-2} - 1| = 0$ and the required bound trivially holds). Then

$$\begin{aligned}
\|D_Y^2 \hat{D}_A^{-2} - I\|_\infty &= \max_{i=1,\ldots,n} \left|D_{Y,ii}^2 \hat{D}_{A,ii}^{-2} - 1\right| \\
&= \max_{i=1,\ldots,n} \left|\left(D_{Y,ii}^2/D_{\beta_0,ii}^2 - 1 + 1\right)\left(D_{\beta_0,ii}^2/\hat{D}_{A,ii}^2 - 1 + 1\right) - 1\right|, \\
&= \max_{i=1,\ldots,n} \left|(\eta_{i,1} + 1)(\eta_{i,2} + 1) - 1\right|, \\
&= \max_{i=1,\ldots,n} \left|\eta_{i,1}\eta_{i,2} + \eta_{i,1} + \eta_{i,2}\right|, \\
&\leq 4d_{\min}^{-2} LK_X s\lambda_A + (2d_{\min}^{-2} LK_X s\lambda_A)^2.
\end{aligned}$$

Finally, by assumption $6d_{\min}^{-2} LK_X s\lambda_A \leq 1/2$ and from the last display and Lemma 2, it follows that

$$|\sigma - 1| \leq \|D_Y^2 \hat{D}_A^{-2} - I\|_\infty \leq 6d_{\min}^{-2} LK_X s\lambda_A =: r_{\text{rem}_0}.$$

We also see from this calculation that under our conditions, $\sigma \geq 1/2$ and $\|D_Y \hat{D}_A^{-1}\|_\infty \leq 2$.

Bound for rem_1**.** A Taylor expansion of μ yields

$$\mu(x_i^T \beta_0) - \mu(x_i^T \hat{\beta}) = \mu'(x_i^T \tilde{\beta}_{(i)}) x_i^T(\beta_0 - \hat{\beta}),$$

where $\tilde{\beta}_{(i)} = \alpha_i \beta_0 + (1 - \alpha_i)\hat{\beta}$ for some $\alpha_i \in [0, 1]$. Let $\tilde{\beta}$ denote a vector with entries $\tilde{\beta}_{(i)}$. Then

$$\text{rem}_1 = \hat{w}_A^T \hat{D}_A^{-1}\{\mu(X\beta_0) - \mu(X\hat{\beta})\} = \hat{w}_A^T \hat{D}_A^{-1} \mu'(X\tilde{\beta}) X(\beta_0 - \hat{\beta}) =: \text{rem}_{11} + \text{rem}_{12},$$

where

$$\text{rem}_{11} = \hat{w}_A^T \hat{D}_A^{-1} \mu'(X\hat{\beta}_A) X(\beta_0 - \hat{\beta}) = \hat{w}_A^T \hat{\Omega}_A X(\beta_0 - \hat{\beta}),$$

and

$$\text{rem}_{12} = \hat{w}_A^T \hat{D}_A^{-1} (\mu'(X\tilde{\beta}) - \mu'(X\hat{\beta}_A)) X(\beta_0 - \hat{\beta}).$$

Using Hölder's inequality together with $\hat{\beta} \in \Theta(\lambda, \beta_0, X)$ and the KKT conditions of the square-root Lasso (5), we have

$$|\text{rem}_{11}| = |\hat{w}_A^T \hat{\Omega}_A X(\beta_0 - \hat{\beta})| \leq \|\hat{w}_A^T \hat{\Omega}_A X\|_\infty \|\beta_0 - \hat{\beta}\|_1 \leq \lambda_{\text{sq}} \sqrt{n} s \lambda.$$

We proceed to bound the second term, rem_{12}. First note that

$$\left|x_i^T \tilde{\beta}_{(i)}\right| \leq \left|x_i^T \beta_0\right| + \left|x_i^T (\tilde{\beta}_{(i)} - \beta_0)\right| \leq K_0/2 + K_0/2 \leq K_0,$$

Then by the Lipschitz property of μ' on $[-K_0, K_0]$ we have

$$|\mu'(x_i^T \tilde{\beta}_{(i)}) - \mu'(x_i^T \hat{\beta}_A)| \leq L|x_i^T (\tilde{\beta}_{(i)} - \hat{\beta}_A)| \leq L\big(|x_i^T (\hat{\beta} - \beta_0)| + |x_i^T (\beta_0 - \hat{\beta}_A)|\big), \tag{43}$$

Then, on the event that $\hat{\beta} \in \Theta(\lambda, \beta_0, X)$,

$$\begin{aligned}
|\text{rem}_{12}| &:= \left|\hat{w}_A^T \hat{D}_A^{-1} (\mu'(X\tilde{\beta}) - \mu'(X\hat{\beta}_A)) X(\beta_0 - \hat{\beta})\right| \\
&= \left|\sum_{i=1}^n \hat{w}_{A,i} \hat{D}_{A,ii}^{-1} \left\{\mu'(x_i^T \tilde{\beta}_{(i)}) - \mu'(x_i^T \hat{\beta}_A)\right\} x_i^T (\beta_0 - \hat{\beta})\right| \\
&\leq \sum_{i=1}^n \left|\hat{w}_{A,i} \hat{D}_{A,ii}^{-1}\right| \left|\mu'(x_i^T \tilde{\beta}_{(i)}) - \mu'(x_i^T \hat{\beta}_A)\right| \left|x_i^T (\beta_0 - \hat{\beta})\right| \\
&\overset{(43)}{\leq} L \sum_{i=1}^n \left|\hat{w}_{A,i} \hat{D}_{A,ii}^{-1}\right| \left(\frac{3}{2}|x_i^T (\beta_0 - \hat{\beta})|^2 + \frac{1}{2}|x_i^T (\beta_0 - \hat{\beta}_A)|^2\right) \\
&\leq L \max_{i=1,\ldots,n} \left|\hat{w}_{A,i} \hat{D}_{A,ii}^{-1}\right| \sum_{i=1}^n \left(\frac{3}{2}|x_i^T (\beta_0 - \hat{\beta})|^2 + \frac{1}{2}|x_i^T (\beta_0 - \hat{\beta}_A)|^2\right) \\
&\leq 4L d_{\min}^{-1} \|\hat{w}_A\|_\infty s(\lambda^2 + \lambda_A^2) n.
\end{aligned}$$

Therefore, $\mathbb{P}_A(|\text{rem}_1| \geq r_{\text{rem}_1}) \leq \delta$, where

$$r_{\text{rem}_1} := \lambda_{\text{sq}} \sqrt{n} s \lambda + 4L d_{\min}^{-1} \|\hat{w}_A\|_\infty s(\lambda^2 + \lambda_A^2) n.$$

We conclude, using (40) and (41), and taking $\epsilon := r_{\text{rem}_1}$ that

$$\begin{aligned}
|\mathbb{P}_A\left(T - \boldsymbol{\Delta} < z\right) - \Phi(z)| &\leq \frac{C_1 C_\varepsilon}{\sigma} \|D_Y \hat{D}_A^{-1} \hat{w}_A\|_\infty + \frac{\sqrt{2} r_{\text{rem}_1}}{\sqrt{\pi}} + \delta + \frac{\sqrt{2} r_{\text{rem}_0}}{\sqrt{\pi}} \\
&\leq 4 C_1 C_\varepsilon \|\hat{w}_A\|_\infty + \frac{\sqrt{2} r_{\text{rem}_1}}{\sqrt{\pi}} + \delta + \frac{\sqrt{2} r_{\text{rem}_0}}{\sqrt{\pi}}.
\end{aligned}$$

The result follows. □

6.4 Proofs for Section 3.3

The logistic loss function is

$$\rho(u, y) := -yu + d(u),$$

where $d(\xi) := \log\big(1 + e^{\xi}\big)$, and we let $f_\beta(x) := x^T\beta$. We define the risk function

$$R(f|X) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big\{\rho\big(f(x_i), Y_i\big) \mid X\big\}$$

and set $\beta_0 := \operatorname{argmin}_{\beta\in\mathbb{R}^p} R(f_\beta|X)$.

Proof of Corollary 1. We apply Theorem 3 to the case of logistic regression to obtain local guarantees on the power of the test. To this end, we need to bound δ in (11) and Condition 3 of Theorem 3.

To bound δ, we note that by Lemma 6 in Section A with $t := \log(2p)$, we have with probability at least $1 - 1/(2p)$ that

$$R\big(f_{\hat\beta}|X\big) - R\big(f_0|X\big) + \lambda\|\hat\beta - \beta_0\|_1 \le \frac{17\lambda^2 s(e^\eta/\epsilon_0 + 1)^2}{\phi^2}.$$

In what follows we work on the event where this occurs. We next want to obtain a bound on $\|X(\hat\beta - \beta_0)\|_2^2/n$. Note that the second derivative of the loss function is

$$\frac{\partial^2\rho(u, y)}{\partial u^2} = d''(u) = \frac{e^u}{1 + e^u}\left(1 - \frac{e^u}{1 + e^u}\right).$$

For $\|x\|_\infty \le K$ and any f with $\sup_{x:\|x\|_\infty\le K} |f(x) - f_0(x)| \le \eta$, we therefore have

$$d''(f(x)) \ge (e^{|f_0(x)|+\eta} + 1)^{-2} \ge (e^\eta/\epsilon_0 + 1)^{-2} =: c_0^2 > 0. \tag{44}$$

Note that for any $\tilde\beta$ on the line segment between β_0 and $\hat\beta$, we have

$$\begin{aligned}\sup_{x:\|x\|_\infty\le K} |f_{\tilde\beta}(x) - f_0(x)| &\le \sup_{x:\|x\|_\infty\le K} |f_{\tilde\beta}(x) - f_{\beta_0}(x)| + \eta/2 \\ &\le K\|\hat\beta - \beta_0\|_1 + \eta/2 \le \eta.\end{aligned}$$

Thus we can conclude using a Taylor expansion of the loss function that there exist $\{\tilde\beta_{(i)} : i = 1, \dots, n\}$, each on the line segment from β_0 to $\hat\beta$, such that

$$\begin{aligned}R(f_{\hat\beta}|X) - R(f_{\beta_0}|X) &= \frac{1}{2n}\sum_{i=1}^{n} d''(x_i^T\tilde\beta_{(i)})\big(x_i^T(\hat\beta - \beta_0)\big)^2 \\ &\ge c_0^2\|X(\hat\beta - \beta_0)\|_2^2/(2n).\end{aligned}$$

We deduce that there exists a constant $\tilde C > 0$ such that with $\lambda = \tilde C\sqrt{\log(2p)/n}$, we have $\delta = \mathbb{P}\big(\hat\beta \notin \Theta(\lambda, \beta_0, X)|X\big) \le 1/(2p)$.

It remains to check that Condition 3 of Theorem 3 is satisfied. Firstly, the inverse link function $\mu(u) = 1/(1 + e^{-u})$ is differentiable and Lipschitz with constant 1. Moreover, by (44), $D_{Y,ii}^2 \ge d_{\min}^2 := c_0^2$ and also $D_{\beta_0,ii}^2 \ge c_0^2$. Finally, observe that $\mathbb{E}\big\{|Y_i - \mu(f_0(x_i))|^3|X\big\} \le 1$. Moreover, $12d_{\min}^{-2}LK_X s\lambda = 12c_0^{-2}LK_X s\lambda \le 1$ by hypothesis. Therefore, Condition 3 is satisfied. □

Acknowledgements: The authors would like to thank the Isaac Newton Institute for Mathematical Sciences for support and hospitality during the 'Statistical Scalability' programme when work on this paper was undertaken, supported by EPSRC grant number EP/R014604/1. JJ is supported by a Swiss National Science Foundation fellowship. RDS is supported by an EPSRC First Grant and an EPSRC programme grant. PB is supported by the European Research Council under the grant agreement No. 786461 (CausalStats – ERC-2017-ADG). RJS is supported by an EPSRC fellowship and an EPSRC programme grant.

References

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-root Lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer–Verlag, Berlin.

Cai, T. T. and Guo, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Annals of Statistics*, 45(2): 615 – 646.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Annals of Statistics*, 41(6):2786–2819.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics*, 7(1):649–688.

Chetverikov, D. and Chernozhukov, V. (2016). On cross-validated Lasso. *arXiv preprint arXiv:1605.02214*.

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015). High-dimensional inference: Confidence intervals, p-values and R-software hdi. *Statistical Science*, 30(4):533–558.

Dezeure, R., Bühlmann, P., and Zhang, C.-H. (2017). High-dimensional simultaneous inference with the bootstrap. *TEST*, 26(4):685–719.

Esseen, C.-G. (1942). On the Liapunov limit error in the theory of probability. *Ark. Mat. Astr. Fys.*, 28:1–19.

Farrington, C. (1996). On assessing goodness of fit of generalized linear models to sparse data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:349–360.

Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference*.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Goeman, J. J., van De Geer, S. A., De Kort, F., and Van Houwelingen, H. C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99.

Hosmer, D. W. and Hjort, N. L. (2002). Goodness-of-fit processes for logistic regression: simulation results. *Statistics in Medicine*, 21(18):2723–2738.

Hosmer, D. W., Hosmer, T., Le Cessie, S., and Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16(9):965–980.

Hosmer, D. W. and Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics*, 80:1043–1069.

Janková, J. and Shah, R. D. and Bühlmann, P. and Samworth, R. J. (2019). GRPtests: Goodness-of-Fit Tests in High-Dimensional GLMs. *R package version 0.1.0. Available at CRAN* https://cran.r-project.org/web/packages/GRPtests/index.html.

Javanmard, A. and Lee, J. D. (2017). A flexible framework for hypothesis testing in high-dimensions. *arXiv preprint arXiv:1704.07971*.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909.

Kohl, M. (2018). *MKmisc: Miscellaneous functions from M. Kohl*. R package version 1.2.

Le Cessie, S. and Van Houwelingen, J. (1991). A goodness-of-fit test for binary regression models, based on smoothing methods. *Biometrics*, 47:1267–1282.

Lemeshow, S. and Hosmer Jr, D. W. (1982). A review of goodness of fit statistics for use in the development of logistic regression models. *American Journal of Epidemiology*, 115(1):92–106.

Lin, D., Wei, L., and Ying, Z. (2002). Model-checking techniques based on cumulative residuals. *Biometrics*, 58(1):1–12.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, volume 37. CRC Press.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). p-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681.

Ning, Y., Zhao, T., and Liu, H. (2017). A likelihood ratio framework for high dimensional semiparametric regression. *Annals of Statistics*, 45(4):2299–2327.

Osius, G. and Rojek, D. (1992). Normal goodness-of-fit tests for multinomial models with large degrees of freedom. *Journal of the American Statistical Association*, 87(420):1145–1152.

Park, M. Y. and Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677.

Shah, R. D. and Bühlmann, P. (2018). Goodness-of-fit tests for high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):113–135.

Shah, R. D. and Samworth, R. J. (2013). Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):55–80.

Spira, A., Beane, J., Shah, V., and Steiling, K. (2007). Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nature Medicine*, 13(3):361–366.

Su, J. Q. and Wei, L. (1991). A lack-of-fit test for the mean function in a generalized linear model. *Journal of the American Statistical Association*, 86(414):420–426.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika*, 99(4):879–898.

Tsiatis, A. A. (1980). A note on a goodness-of-fit test for the logistic regression model. *Biometrika*, 67(1):250–251.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of Statistics*, 37(5):2178–2201.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics*, 14(4):1261–1295.

Yu, Y., Bradic, J., and Samworth, R. J. (2020). Confidence intervals for high-dimensional Cox models. *Statistica Sinica*, to appear.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low-dimensional parameters in high-dimensional linear models. *Journal of the Royal Statistical Society: Series B*, 76:217–242.

Zhu, Y. and Bradic, J. (2017). A projection pursuit framework for testing general high-dimensional hypothesis. *arXiv preprint arXiv:1705.01024*.

A Online Supplementary Material for "Goodness-of-fit testing in high-dimensional generalized linear models"

This appendix contains technical results and proofs omitted in the main paper.

A.1 Auxiliary lemmas

Lemma 1 (Hoeffding's inequality for a maximum of p averages)**.** *Suppose that for each $j = 1, \ldots, p$, the random variables $Z_{1j}, \ldots, Z_{nj}$ are independent with*

$$\mathbb{E}Z_{ij} = 0, \quad |Z_{ij}| \leq c_i.$$

Then for all $t > 0$

$$\mathbb{P}\left(\max_{j=1,\ldots,p} \left| \frac{1}{n} \sum_{i=1}^{n} Z_{ij} \right|^2 \geq \frac{\|c\|_2^2}{n} \frac{2(t + \log(2p))}{n} \right) \leq e^{-t}.$$

Proof of Lemma 1. Apply Corollary 17.1 in van de Geer (2016) together with a union bound. □

Lemma 2. *Let $\tilde{A}, \tilde{B} \in \mathbb{R}^{n \times n}$ be diagonal matrices and suppose that $\tilde{B}$ is invertible. Let $w \in \mathbb{R}^n$ satisfy $\|w\|_2 = 1$. Then*

$$\left| \frac{\|\tilde{A}w\|_2}{\|\tilde{B}w\|_2} - 1 \right| \leq \|\tilde{A}^2\tilde{B}^{-2} - I\|_\infty.$$

Proof of Lemma 2.

$$|\|\tilde{A}w\|_2^2 - \|\tilde{B}w\|_2^2| = |w^T(\tilde{A}^2 - \tilde{B}^2)w| \leq \max_{i=1,\ldots,n} \frac{|\tilde{A}_{ii}^2 - \tilde{B}_{ii}^2|}{\tilde{B}_{ii}^2} \|\tilde{B}w\|_2^2 = \|\tilde{A}^2\tilde{B}^{-2} - I\|_\infty \|\tilde{B}w\|_2^2.$$

Hence

$$\left| \frac{\|\tilde{A}w\|_2}{\|\tilde{B}w\|_2} - 1 \right| = \frac{\left| \|\tilde{A}w\|_2^2 / \|\tilde{B}w\|_2^2 - 1 \right|}{\left| \|\tilde{A}w\|_2 / \|\tilde{B}w\|_2 + 1 \right|} \leq \|\tilde{A}^2\tilde{B}^{-2} - I\|_\infty,$$

as required. □

A.2 Auxiliary lemmas for Group Testing

Lemma 3. *Under the conditions of Proposition 1, we have*

$$\mathbb{P}(\mathcal{T}^c) \leq 3/p.$$

Proof. To obtain a probability bound for $\mathcal{T}_{1,j}$, we can apply Lemma 1, noting that

$$\frac{1}{n}\|u_j^T D_{\beta_0} X_{-G}\|_\infty = \max_{k \in G^c} \left| \frac{1}{n} \sum_{i=1}^{n} Z_{ijk} \right|,$$

with $Z_{ijk} := u_{j,i} D_{\beta_0,ii} X_{-G,i,k}$, where $X_{-G,i,k}$ is the (i,k)-th entry of the matrix X_{-G} and $u_{j,i}$ is the i-th entry of u_j. Note that by Condition 2 (ii), it follows that $|u_{j,i}| \leq 2K$ and by Condition 2

(iii), we have $|D_{\beta_0,ii}| \leq C_0$. Therefore, $|Z_{ijk}| \leq c_i$ for $c_i := 2C_0K^2$ and for all i, j, k. Thus Lemma 1 implies that for all $t > 0$,

$$\mathbb{P}\left(\frac{1}{n^2}\|u_j^T D_{\beta_0} X_{-G}\|_\infty^2 \geq 2(2C_0K^2)^2 \frac{t + \log(2p)}{n}\right) \leq e^{-t}.$$

Therefore,

$$\mathbb{P}(\mathcal{T}_{1,j}^c) \leq 1/(2p)^2. \tag{45}$$

For the set $\mathcal{T}_{3,j}$, by the sub-Gaussianity of $\eta_i = Y_i - \mu(x_i^T\beta_0)$ from Condition 2 (i), there exists a constant $C > 0$ such that

$$\mathbb{P}\left(\mathcal{T}_{3,j}^c\right) = \mathbb{P}\left(\max_{k \in G^c}\left|\frac{1}{n}\sum_{i=1}^n X_{j \cup G^c, i, k}(Y_i - \mu(x_i^T\beta_0))\right| \geq CK\sqrt{\frac{\log(2p)}{n}}\right) \leq 1/p^2. \tag{46}$$

Therefore,

$$\mathbb{P}\left(\mathcal{T}_j^c\right) \leq \mathbb{P}(\mathcal{T}_{1,j}^c) + \mathbb{P}(\mathcal{T}_2^c) + \mathbb{P}(\mathcal{T}_{3,j}) \leq 1/p^2 + \mathbb{P}(\mathcal{T}_2^c) + 1/p^2.$$

Using bounds (45) and (46), the fact that $|G| \leq p$ and the assumption $\mathbb{P}(\mathcal{T}_2^c) \leq 1/p$, we obtain by a union bound that

$$\mathbb{P}\left(\cup_{j \in G}\mathcal{T}_j^c\right) \leq |G| \max_{j \in G}\mathbb{P}(\mathcal{T}_{1,j}^c) + \mathbb{P}(\mathcal{T}_2^c) + |G| \max_{j \in G}\mathbb{P}(\mathcal{T}_{3,j}^c) \leq 2/p + \mathbb{P}(\mathcal{T}_2^c) \leq 3/p.$$

□

Lemma 4. *Assume Conditions 1 and 2. For $j \in G$, we let $u_j := X_j - X_{-j}\gamma_{0,j}$ and define the sets*

$$\mathcal{T}_{1,j} := \left\{\|u_j^T D_{\beta_0} X_{-G}\|_\infty/n \leq 6C_0K^2\sqrt{\log(2p)/n}\right\},$$

$$\mathcal{T}_2 := \left\{\hat{\beta}_{-G} \in \Theta_{-G}(\lambda, \beta_0)\right\}.$$

Then there exist $\lambda_{\text{nw}} \asymp \sqrt{\log p/n}$, and positive constants C_3 and C_4 such that on the set $\cap_{j \in G}(\mathcal{T}_{1,j} \cap \mathcal{T}_2)$, it holds that

$$\max_{j \in G}\|\hat{\gamma}_j - \gamma_{0,j}\|_1 \leq C_3K^2 s\sqrt{\log p/n},$$
$$\max_{j \in G}|\hat{\tau}_j^2 - \tau_j^2| \leq C_4K^2\sqrt{s \log p/n},$$

whenever $\hat{\tau}_j^2 := \|\hat{D}(X_j - X_{-G}\hat{\gamma}_j)\|_2^2/n > 0$.

Proof of Lemma 4. To obtain rates of convergence for $\hat{\gamma}_j$ from Algorithm 2, we follow the arguments in the proof of Theorem 3.2 in van de Geer et al. (2014), which considers nodewise regression with random bounded design. The difference is that they define a nodewise regression program with design matrix X_{-j} and use the Lasso, whereas we want to apply the nodewise regression with a smaller design matrix, X_{-G}, and we use the square-root Lasso. We also seek finite-sample, as opposed to asymptotic, bounds, but this requires only minor modifications. But since we assume that $\hat{\tau}_j > 0$, the square-root Lasso program with penalty λ_{nw} corresponds to the Lasso program with

penalty $\lambda_{\text{Lasso}} := \hat{\tau}_j \lambda_{\text{nw}}$. We now check that the appropriate finite-sample analogues of conditions (D1)–(D5) of Theorem 3.2 from van de Geer et al. (2014) are satisfied for $\tilde{X} := X_{j \cup G^c}$. Firstly, the analogues of (D1), (D2), (D4) are satisfied directly by the assumptions in Conditions 1 and 2. For (D3), first note that the smallest eigenvalue of $\Sigma_{\beta_0, j \cup G^c} := \mathbb{E}\tilde{X}^T D^2_{\beta_0} \tilde{X}/n$ is lower bounded by the smallest eigenvalue of $\Sigma_0 = \mathbb{E}X^T D^2_{\beta_0} X/n$, which is in turn lower bounded by $1/C_e$. Similarly, $\|\Sigma_{\beta_0, j \cup G^c}\|_\infty \leq \|\Sigma_{\beta_0}\|_\infty \leq C_e$,. Finally, Condition (D5) is satisfied on $\mathcal{T}_2$.

As in the proof of Proposition 1, let k_j denote the index of column X_j in the matrix $X_{j \cup G^c}$. We write

$$\hat{\Gamma}_j := (-\hat{\gamma}_{j,1}, \ldots, -\hat{\gamma}_{j,k_j-1}, 1, -\hat{\gamma}_{j,k_j+1}, \ldots, -\hat{\gamma}_{j,|G^c|})^T,$$

and $\Gamma_{0,j}$ for its analogy defined in terms of $\gamma_{0,j}$. Then note that we can write $X_j - X_{-G}\hat{\gamma}_j = \tilde{X}\hat{\Gamma}_j$ and recall that $\hat{\tau}_j^2 = \|\hat{D}\tilde{X}\hat{\Gamma}_j\|_2^2/n$ and $\tau_j^2 = \mathbb{E}\|D_{\beta_0}\tilde{X}\Gamma_{0,j}\|_2^2/n$. By inspecting the proof of Theorem 3.2 of van de Geer et al. (2014), we conclude that there exist positive constants C_3 and C_4 such that on $\cap_{j \in G}(\mathcal{T}_{1,j} \cap \mathcal{T}_2)$, it holds that

$$\max_{j \in G} \|\hat{\gamma}_j - \gamma_{0,j}\|_1 \leq C_3 K^2 s \sqrt{\log p/n},$$
$$\max_{j \in G} |\hat{\tau}_j^2 - \tau_j^2| \leq C_4 K^2 \sqrt{s \log p/n},$$

as required. □

The following lemma bounds a term Δ_2 defined in the proof of Theorem 2.

Lemma 5. *Under the conditions of Theorem 2, there exists a constant $C'' > 0$ such that*

$$\mathbb{P}\big(\Delta_2 \geq (C'')^2 K^6 s^2 \lambda^2\big) \leq 4/p.$$

Proof of Lemma 5. On the set $\mathcal{T}$ defined in the proof of Proposition 1, we have

$$\begin{aligned}
\Delta_2 &= \max_{j \in H} \sigma_j^2 = \max_{j \in H} \frac{1}{n} \sum_{i=1}^n (x_{ij} - \hat{x}_{ij})^2 \\
&= \max_{j \in H} \sum_{i=1}^n (w_{j,i}\varepsilon_i - \hat{w}_{j,i}\hat{R}_{G,i})^2 \qquad (47) \\
&\leq 2 \max_{j \in H} \sum_{i=1}^n (w_{j,i} - \hat{w}_{j,i})^2 \varepsilon_i^2 + 2 \max_{j \in H} \sum_{i=1}^n \hat{w}_{j,i}^2 (\hat{R}_{G,i} - \varepsilon_i)^2 \\
&=: r_1 + r_2. \qquad (48)
\end{aligned}$$

Now we bound r_1. By similar arguments as in the proof of Proposition 1, we will now show that on $\mathcal{T}$,

$$\max_{i,j} (w_{j,i} - \hat{w}_{j,i})^2 = \mathcal{O}(K^3 s^2 \lambda^2/n).$$

First,

$$\begin{aligned}\max_{i,j}|w_{j,i}-\hat{w}_{j,i}| &= \max_{i,j}\frac{1}{\sqrt{n}}|D_{\beta_0,ii}e_i^T X_{j\cup G^c}\Gamma_{0,j}/\tau_j - D_{\hat{\beta}_{-G},ii}e_i^T X_{j\cup G^c}\hat{\Gamma}_j/\hat{\tau}_j| \\ &\le \max_{i,j}\frac{1}{\sqrt{n}}|(D_{\beta_0,ii}-D_{\hat{\beta}_{-G},ii})e_i^T X_{j\cup G^c}\Gamma_{0,j}/\tau_j| \\ &\quad + \frac{1}{\sqrt{n}}\max_{i,j}|D_{\hat{\beta}_{-G},ii}e_i^T X_{j\cup G^c}(\hat{\Gamma}_j/\hat{\tau}_j - \Gamma_{0,j}/\tau_j)|.\end{aligned}$$

Now we can use Condition 2(ii), (19), (28), (29) and (24) to bound the terms in the last display and obtain

$$\begin{aligned}\max_{i,j}|w_{j,i}-\hat{w}_{j,i}| &\le \frac{1}{\sqrt{n}}\max_i\big|D_{\beta_0,ii}-D_{\hat{\beta}_{-G},ii}\big|\max_{i,j}|e_i^T X_{j\cup G^c}\Gamma_{0,j}/\tau_j| \\ &\quad + \frac{1}{\sqrt{n}}\max_{i,j}\big|D_{\hat{\beta}_{-G},ii}\big|\|X_{j\cup G^c}\|_\infty\|\hat{\Gamma}_j/\hat{\tau}_j - \Gamma_{0,j}/\tau_j\|_1 \\ &\lesssim \frac{1}{\sqrt{n}}LK^2 s\lambda + \frac{1}{\sqrt{n}}K^3 s\sqrt{\log p/n} \\ &\lesssim \frac{1}{\sqrt{n}}K^3 s\lambda.\end{aligned}$$

Moreover, by the sub-Gaussianity of η_i and since $D^2_{\beta_0,ii} \ge c_0$ (by Condition (2)(iii)), there exist constants $C, C''' > 0$ such that with probability at least $1-1/p$, we have

$$\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2 \le \frac{1}{n}\sum_{i=1}^n \mathbb{E}\varepsilon_i^2 + C\sqrt{\frac{\log p}{n}} \le C'''. \tag{49}$$

Therefore, with probability at least $1-1/p-\mathbb{P}(\mathcal{T}^c)$,

$$r_1 = \max_j \sum_{i=1}^n (w_{j,i}-\hat{w}_{j,i})^2\varepsilon_i^2 \le \max_{i,j}(w_{j,i}-\hat{w}_{j,i})^2\sum_{i=1}^n\varepsilon_i^2 \lesssim K^6 s^2\lambda^2. \tag{50}$$

We now bound r_2. Using Condition 2(iii), together with the fact that $\hat{\beta}_{-G} \in \Theta_{-G}(\lambda,\beta_0)$ on $\mathcal{T}$ we have that on this event,

$$\begin{aligned}|\hat{R}_{G,i}-\varepsilon_i| &= \big|D^{-1}_{\hat{\beta}_{-G},ii}\big(Y_i-\mu(X_{i,-G}\hat{\beta}_{-G})\big) - D^{-1}_{\beta_0,ii}\big(Y_i-\mu(X_{i,-G}\beta_{0,-G})\big)\big| \\ &\le C_0\big|D^{-1}_{\hat{\beta}_{-G},ii}X_{i,-G}(\beta_{0,-G}-\hat{\beta}_{-G})\big| + \big|D^{-1}_{\hat{\beta}_{-G},ii} - D^{-1}_{\beta_0,ii}\big|\big|Y_i-\mu(X_{i,-G}\beta_{0,-G})\big| \\ &\le C_0\big|D^{-1}_{\hat{\beta}_{-G},ii}X_{i,-G}(\beta_{0,-G}-\hat{\beta}_{-G})\big| + D^{-1}_{\beta_0,ii}\left|\frac{D_{\beta_0,ii}}{D_{\hat{\beta}_{-G},ii}}-1\right||\eta_i|.\end{aligned}$$

Then using the fact that $\hat{\beta}_{-G} \in \Theta_{-G}(\lambda,\beta_0)$ on $\mathcal{T}$ and using that

$$\left|\frac{D_{\beta_0,ii}}{D_{\hat{\beta}_{-G},ii}}-1\right| \le \left|\frac{D^2_{\beta_0,ii}}{D^2_{\hat{\beta}_{-G},ii}}-1\right| \le 2c_0^{-1}LKs\lambda$$

(which follows similarly as in the proof of Theorem 3) and using (49), we obtain that on $\mathcal{T}$,

$$\sum_{i=1}^{n}(\hat{R}_{G,i}-\varepsilon_i)^2 \le 2C_0\sum_{i=1}^{n} D^{-2}_{\hat{\beta}_{-G},ii}|X_{i,-G}(\beta_{0,-G}-\hat{\beta}_{-G})|^2 + 2\sum_{i=1}^{n}\left|\frac{D_{\beta_0,ii}}{D_{\hat{\beta}_{-G},ii}}-1\right|^2\varepsilon_i^2$$
$$\lesssim s\lambda^2 n + K^2 s^2\lambda^2 n \lesssim K^2 s^2\lambda^2 n.$$

Then

$$r_2 = 2\max_{j\in H}\sum_{i=1}^{n}\hat{w}_{j,i}^2(\hat{R}_{G,i}-\varepsilon_i)^2 \lesssim \frac{K^2}{n}K^2 s^2\lambda^2 n = K^4 s^2\lambda^2, \tag{51}$$

where we used $\|\hat{w}_j\|_\infty^2 = \mathcal{O}(K^2/n)$ (which follows from (30)).

Overall, collecting (51) and (50) there exists a constant C'' such that

$$\mathbb{P}(\Delta_2 \ge (C'')^2 K^6 s^2\lambda^2) \le 1/p + \mathbb{P}(\mathcal{T}^c) \le 4/p. \tag{52}$$

□

A.3 Multiplier bootstrap

We summarize Corollary 3.1 from Chernozhukov et al. (2013). To this end, we need the following condition.

Condition 4. *Let $(x_i)_{i=1}^n$ be n independent random vectors with values in $\mathbb{R}^g$ satisfying*

$$c_1 \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}x_{ij}^2 \le C_1 \tag{53}$$

and

$$\max_{k=1,2}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|x_{ij}|^{2+k}/B_n^k + \mathbb{E}\max_{1\le j\le g}|x_{ij}/B_n|^4 \le 4. \tag{54}$$

Define

$$T_0 := \max_{1\le j\le g}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_{ij}.$$

Let $(e_i)_{i=1}^n$ be a sequence of i.i.d. $\mathcal{N}(0,1)$ random variables independent of $(x_i)_{i=1}^n$ and define

$$W_0 := \max_{1\le j\le g}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_{ij}e_i.$$

Assume that there exist $\zeta_1,\zeta_2 \ge 0$ such that

$$\mathbb{P}(|T-T_0|>\zeta_1)\le\zeta_2, \qquad \mathbb{P}(\mathbb{P}_e(|W-W_0|>\zeta_1)>\zeta_2)<\zeta_2, \tag{55}$$

where $\mathbb{P}_e$ is the probability measure induced by the multiplier variables $(e_i)_{i=1}^n$ holding $(x_i)_{i=1}^n$ fixed.

Theorem 4 (Corollary 3.1 in Chernozhukov et al. (2013)). *Suppose that Condition 4 is satisfied with*

$$\zeta_1 \sqrt{\log g} + \zeta_2 \leq C_2 n^{-c_2} \tag{56}$$

and

$$B_n^4 \log(gn)^7 / n \leq C_2 n^{-c_2}. \tag{57}$$

Then there exist constants $c, C > 0$ depending only on c_1, C_1, c_2 and C_2 such that

$$\sup_{\alpha \in (0,1)} |\mathbb{P}(T < c_W(\alpha)) - \alpha| \leq C n^{-c},$$

where $c_W(\alpha)$ is the α-quantile of W conditional on $(x_i)_{i=1}^n$ given by

$$c_W(\alpha) := \inf\{t \in \mathbb{R} : \mathbb{P}_e(W \leq t) \geq \alpha\}.$$

A.4 Oracle inequalities for logistic regression under misspecification

We require a condition on the design matrix known as the compatibility condition (Bühlmann and van de Geer, 2011).

Definition 1 (Compatibility constant). We say that the compatibility condition is met with constant $\phi > 0$ if for all $\beta \in \mathbb{R}^p$ that satisfy $\|\beta_{S^c}\|_1 \leq 3\|\beta_S\|_1$, it holds that

$$\|\beta_S\|_1^2 \leq \frac{s\|X\beta\|_2^2}{\phi^2}.$$

For our final lemma, we use the notation of Sections 3.3 and 6.4.

Lemma 6. *Suppose that there exists a constant $K > 0$, such that*

$$\max_{1 \leq j \leq p} \|X_j\|_\infty \leq K, \quad \max_{1 \leq j \leq p} \|X_j\|_2 \leq 1,$$

and that X satisfies the compatibility condition with constant $\phi > 0$. Take $t > 0$ and let

$$\bar{\lambda} := \sqrt{2\log(2p)/n} + K\log(2p)/(3n),$$

$$\lambda_0 := 4\bar{\lambda} + tK/(3n) + \sqrt{2t(1 + 8\bar{\lambda})/n}.$$

Assume that there exist constants ϵ_0, η such that

$$0 < \eta \leq \epsilon_0 < \pi_0(x) < 1 - \epsilon_0, \quad \textit{for all } \|x\|_\infty \leq K,$$

$$\sup_{x : \|x\|_\infty \leq K} |f_{\beta_0}(x) - f_0(x)| \leq \eta/2.$$

For some constant $M \geq 8$ take λ satisfying $8\lambda_0 \leq \lambda \leq M\lambda_0$ and $17\lambda s(e^\eta/\epsilon_0 + 1)^2/\phi^2 \leq \eta/(2K)$, and further assume that

$$R(f_{\beta_0}|X) - R(f_0|X) \leq \min\left\{\eta\lambda_0/4, \frac{\lambda^2 s(e^\eta/\epsilon_0 + 1)^2}{6\phi^2}\right\},$$

$$\frac{8KM^2(e^\eta/\epsilon_0 + 1)^2}{\eta} \frac{\lambda_0 s}{\phi^2} \leq 1.$$

Then with probability at least $1 - e^{-t}$, *it holds that*

$$R(f_{\hat{\beta}}|X) - R(f_0|X) + \lambda\|\hat{\beta} - \beta_0\|_1 \leq \frac{17\lambda^2 s(e^{\eta}/\epsilon_0 + 1)^2}{\phi^2}.$$

Proof of Lemma 6. The proof follows from Lemma 6.8 and Section 6.7 in Bühlmann and van de Geer (2011). □

References

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer–Verlag, Berlin.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Annals of Statistics*, 41(6):2786–2819.

van de Geer, S. (2016). *Estimation and Testing under Sparsity: École d'Été de Saint-Flour XLV*. Springer.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202.